Neural and behavioral investigations of individual differences in social learning and decision making

Dissertation
submitted to the Faculty of Economics,
Business Administration and Information Technology
of the University of Zurich

to obtain the degree of
Doktorin der Neuroökonomie, Dr. sc.
(corresponds to Doctor of Neuroeconomics, PhD)

presented by

Azade Seid-Fatemi
from Germany

approved in February 2015 at the request of
Prof. Dr. Philippe Tobler
Prof. Dr. Todd Hare

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 11.02.2015

Chairman of the Doctoral Board: Prof. Dr. Todd Hare

Acknowledgments

Foremost, I would like to express my sincere appreciation and gratitude to my supervisor Professor Philippe Tobler, who patiently guided and supported me throughout my PhD and allowed me to grow as a research scientist. Thank you for your enthusiasm and for always keeping your door open for scientific discussions. This dissertation would not exist without your valuable contribution and advice.

I also wish to thank Professor Alexander Wagner for the fruitful and pleasant scientific discussions we had and Professor Todd Hare for taking the time to be the co-supervisor of my dissertation.

Moreover, I want to thank Sunhae Sul and Yosuke Morishima for the inspiring collaborative work. Although the research projects we conducted together are not included in this thesis, I would like to express my gratitude for your help and valuable suggestions, and for always having an open ear for my worries.

Further thanks go to my master and bachelor students, who did a great job and helped me with data collection. I also want to thank all my current and former colleagues in the lab and in the department of economics for their help and their friendship. In particular, I am grateful to Friederike Meyer for her steady support. Special thanks also go to Tamara Herz, Sally Gschwend, and Karl Treiber for their administrative and technical support and for always being there when needed. For proof-reading this thesis I thank Nadja Freund and Joswin Kattoor.

Finally, I thank my family and friends for their ongoing support. Most importantly, I would like to give very special thanks to Ürün Dogan for his care and endless encouragement throughout my studies.

Abstract

Recent neuroscientific research has begun to explore the neural mechanisms of social behavior, ranging from more basic processes, such as social learning, to more complex social decisions, such as moral judgments. While these studies provided valuable insights into some of the relevant mechanisms, important issues have remained unaddressed. For instance, some of the formal learning principles have not been explored in the social domain. Further, it is not clear how different brain regions communicate with each other during moral decision making. Most importantly, how can individual differences in behavior and disposition prove useful in addressing these questions? The dissertation consists of three studies that aimed to close these gaps.

The first study investigated whether an efficient learning principle, which applies to individual learning, also applies to social learning. This principle is expressed by the blocking effect that refers to the phenomenon that a novel stimulus is blocked from learning when it is associated with a fully predicted outcome. I employed a social variant of the blocking paradigm and used functional magnetic resonance imaging (fMRI) to address the question whether the blocking effect manifests itself when individuals learn about others' rewards in a social context. I demonstrate that blocking does indeed occur in the social domain and it does so to a similar degree as in the individual domain. On the neural level, individual differences in the degree of social blocking relate specifically to activations in the medial prefrontal cortex (mPFC). Thus, while the same efficiency principle applies to reward learning in the individual and social domain, the mPFC plays a central role in implementing it specifically during social learning.

The second study investigated circumstances that prevent the blocking effect and instead lead to learning. Previous research on individual learning has shown that a change in both the value and the identity of a reward results in unblocking, that is learning, rather than blocking. I examined whether and how this would occur in a social context and employed a social variant of the unblocking experiment to assess learning driven by changes in the reward recipient. I found that participants who normally block redundant learning in the standard blocking paradigm show unblocking when the reward recipient changes. Moreover, the degree of recipient-specific unblocking was higher in less empathic and less prosocial participants. This suggests that value-matched learning is sensitive to the reward recipient and that individual differences in unblocking may relate to social traits.

In the third study I examined how individual differences in moral disposition influence behavior and neural correlates during decisions concerning honesty. Specifically, I used fMRI

to study the influence of “protected values”, which supposedly shield moral decisions against economic considerations, when honesty is economically costly. With increasing economic costs I find that participants show stronger activation in the dorsolateral (dlPFC) and dorsomedial (dmPFC) prefrontal cortex the more honest they are. Importantly, functional connectivity between these regions and the inferior frontal gyrus (IFG) was stronger in high cost conditions compared with low cost conditions in participants with high protected values, suggesting the involvement of control mechanisms. Furthermore, the relation between cost-dependent dlPFC-IFG connectivity and protected values was specific for moral decisions as compared with non-moral decisions. These findings provide novel insights into how prefrontal connectivity predicts individual variability in honesty and stress the importance of investigating individual differences in functional connectivity related to moral decision making.

In summary, the result of the studies presented in this thesis show how individual differences in behavior and disposition can help to achieve a more comprehensive understanding of the behavioral and neural underpinnings in social learning and moral decision making.

List of manuscripts

The dissertation is based on the following research articles:

Study 1:

Seid-Fatemi A, Tobler PN (2014). Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex. Soc Cogn Affect Neurosci. Advance online publication. doi: 10.1093/scan/nsu130.

Study 2:

Seid-Fatemi A, Tobler PN. Social unblocking. *In preparation.*

Study 3:

Seid-Fatemi A, Heise F, Tanner C, Gibson R, Wagner AF, Tobler PN. Prefrontal connectivity predicts individual differences in honesty. *In preparation.*

Contents

1 General Introduction

For humans, as for many other species, effective functioning strongly depends on social cognition: the ability to encode and process information about others and to use these representations to guide behavior in a social environment. The basis of social cognition is social learning, which refers to the acquisition of social information or social behavior. More is known about non-social than social learning, which is surprising given the importance of social learning. I wondered whether social learning relies on similar learning mechanisms as non-social learning. Recent neuroscientific research has assumed so and therefore extended formal models of learning to the social domain to describe the mechanisms and neural underpinnings that underlie social learning. On a more complex level, social cognition encompasses decisions that relate to social norms and have social consequences. Moral decisions represent such a high-level social cognitive process and are based on judgments of the appropriateness of one's behavior within the context of socialized perceptions of right and wrong. As these kinds of decisions regulate crucial aspects of our social life, a clear understanding of the neurobiological processes is of critical importance.

In this dissertation, I present three studies (see appendix) that investigate the behavioral and neural mechanisms of social learning and moral decisions making, particularly with respect to individual differences. Study 1 investigates whether an optimality principle, which applies to individual learning and is expressed by the so-called blocking effect, also applies to social learning and how individual differences in behavior influence the neural representations of this effect. Study 2 examines in what circumstances this principle no longer holds (leading to an unblocking effect) during social learning and how this relates to individual differences in social traits. Study 3 explores how moral decisions concerning honesty are represented in the brain and how connectivity patterns in prefrontal regions change in relation to individual differences in honesty.

In the introductory chapter, I will summarize the theoretical background of reinforcement learning in the individual and social domain, as well previous findings on the neurobiology of learning. Next, I give an overview of moral decision making in the context of honesty and describe previous research on the underlying neural processes. The three studies will be summarized in the second chapter. Finally, a general discussion and conclusion form the third chapter.

1.1 Reward and reinforcement learning

Learning to associate environmental stimuli with beneficial or aversive outcomes is fundamentally important as it allows predicting biologically relevant events. Predictions about future events, in turn, may help the individual to prepare actions before these events actually occur. Such preparatory mechanisms enable rapid execution of appropriate responses, which can provide an adaptive advantage over competitors.

The formation of cue-outcome associations in humans and animals was first studied by Ivan Pavlov in the nineteenth century (Pavlov, 1927). In his classic experiment Pavlov observed that experienced dogs started salivating before food was actually delivered and that this response could be induced when a bell was rung to indicate that the food was ready. Through this observation the principles of classical (or Pavlovian) conditioning emerged. A biologically relevant unconditioned stimulus (US) such as food, elicits an unconditioned response (UR; salivation). When a neutral conditioned stimulus (CS; the bell) repeatedly precedes the US, the CS will eventually also cause the dog to salivate (conditioned response, or CR), suggesting that the animal has formed an association between the presentation of the CS and the delivery of the US. In contrast to classical conditioning, instrumental (or operant) conditioning describes the procedure of how animals learn to perform a behavior that leads to a rewarding outcome. This relationship between behavior and reinforcement was first described by Thorndike (Thorndike, 1911) as “the law of effect”, which states that the probability with which animals show a particular behavior increases with the probability of the behavior leading to a positive reinforcement and decreases with the probability of the behavior leading to a punishment. Thus, both types of learning are driven by the formation of associations between the representations of the cue or a behavioral reaction and the reinforcer.

The mechanisms of reinforcement learning have been formalized using mathematical models. The key idea of these models is the notion of surprise that drives learning, which was first formalized by Bush and Mosteller (1951). This earlier model was later extended by Rescorla and Wagner (1972) and in its core still today represents the most influential theoretical model of associative learning. The Rescorla-Wagner (RW) model describes the learning process in terms of the changes in the strength of association between CS and US. The basic assumption is that the change in associative strength ΔV_t at trial t would be represented by:

$$\Delta V_t = \alpha(\lambda_t - \Sigma V_t)$$

Here, λ is the actual outcome on a given trial, ΣV_t the predicted outcome of all stimuli presented, and α is the learning rate. The change in associative strength is driven by the discrepancy between the expected and actual outcome. This difference is termed the prediction error and is thought to act as a teaching signal that updates the value of the CS. On each trial, the associative strength V_t, reflecting the cumulative information from all previous trials, is updated with the change in associative strength ΔV_t in the current trial:

$$V_{t+1} = V_t + \Delta V_t$$

During learning, when the cue-outcome association is newly introduced, the prediction error is large and the associative strength will increase. Over the course of learning, the prediction error will become smaller, and thus the increase in associative strength will also decrease with each trial. Once the outcome is fully predicted, the prediction error will be zero and the associative strength will stay constant as long as the outcome remains unchanged. This happens when learning is complete and has reached an asymptote. The α term reflects the learning rate and determines how much weight is given to recent experience as captured by the prediction error. Thus, the free parameter characterizes how quickly the asymptote is reached.

Although the RW model explains a variety of behaviorally observed associative learning phenomena, it also has some limitations. For example, it cannot account for within-trial effects such as second-order conditioning or sensitivity to stimulus timing. Therefore, more recently the temporal difference (TD) learning model and Q-learning have been introduced as real-time extensions of the RW model that respectively capture classical and instrumental conditioning. These newer models assume that prediction errors are computed at any given moment, both during and in-between trials rather than just at the end of a trial (Sutton and Barto, 1990; Watkins and Dayan, 1992).

1.2 Blocking and unblocking

Although learning is essential for adaptive and goal-directed behavior, it is also costly and requires the use of limited cognitive resources. These resources could be conserved if we prevented learning whenever little or no new information is available. Thus, it would be more efficient if learning only took place when necessary, that is, when previous learning did not make it superfluous. For example, it is efficient not to learn a second cause when a first cause already fully explains an effect (outcome). Associative learning research calls this the “blocking effect”

(Kamin, 1969) because the previously learned association "blocks" learning about a concurrently appearing novel stimulus. The underlying rationale is that learning to predict an outcome on the basis of a novel stimulus is redundant if we can already do so on the basis of a previously learned stimulus.

For instance, assume that Pavlov's dog has learned the association between the bell and the delivery of food and shows a salivating response when presented with the bell. In the next stage, a light is presented in combination with the bell, and both are followed by food. If the light cue is now presented alone, it would not elicit a salivating response. This can be explained by the fact that the food is already fully predicted by the presentation of the bell. As the light provides only redundant information, learning does not occur. Previous learning about the bell has blocked the formation of an association between the light and the reward. By contrast, if the light would appear concurrently with a stimulus that does not predict the reward, then learning about the light is not redundant and blocking does not occur. Blocking of novel learning has been observed in different species including rats (Kamin, 1969), monkeys (Waelti et al., 2001), and humans (Tobler et al., 2006; Prados, 2011; Eippert et al., 2012), who show inter-individual variability in the degree to which this effect occurs (Tobler et al., 2006; Byrom, 2013).

Blocking can be prevented if the value (but not the identity) or the identity (but not the value) of the outcomes is changed. In order to show this "unblocking effect", a variation of the blocking paradigm is used in which the compound of the first and second cue is followed by a different quantity (Holland, 1984) or different identity (Rescorla, 1999) of the outcome than the quantity or identity predicted by the first cue. For example, imagine that Pavlov's dog has learned the association between the bell and the delivery of two food pellets. If in the next stage, a light is presented in combination with the bell and the compound of the two stimuli is followed by two pellets, the light cue would be blocked from learning, as described above. However, if the bell-and-light compound would be followed by one or three pellets, blocking would be prevented and the light cue would be learned. This is because the light cue provides new, non-redundant information about the quantity or value of the food, which is useful to adapt responding. Thus, in contrast to blocking, changing the value of the outcome results in learning of the second cue due to the discrepancy between expected and obtained value.

By contrast, if the value of two outcomes is held constant and only the identity changes, unblocking does not always take place. Evidence for this has come from several studies (Bakal et al., 1974; Dickinson and Dearing, 1979; Ganesan and Pearce, 1988) that have shown how changing the identity, but not the value of the outcome can result in a similar degree of blocking

as keeping identity unchanged. This is the so-called “transreinforcer blocking” phenomenon. To illustrate this effect with Pavlov’s dog experiment, it would mean that the light cue would be blocked to a similar degree as in the standard blocking paradigm, if its presentation in compound with the bell would be followed by, say, two raisins instead of two pellets, a different kind, but equally preferred food. This suggests that at least during some types of learning only value, as processed by a common motivational system, is relevant for learning, while other features such as identity are abstracted away. Transreinforcer blocking and unblocking have been studied in rats (Burke et al., 2008; McDannald et al., 2011; Steinberg et al., 2013) and in humans during causal learning (Le Pelley et al., 2005).

1.3 Neural representations of reward and reinforcement learning

In the last decades a variety of studies investigated the neurobiological mechanisms underlying reward processing and reinforcement learning. Single-unit recordings from dopaminergic neurons in the monkey ventral tegmental area (VTA) and substantia nigra showed that these neurons increase their firing rate to unpredicted rewards (Romo and Schultz, 1990). Subsequent studies showed that if the reward was consistently preceded by a cue, after a number of trials the dopaminergic response in the VTA and substantia nigra shifted from the time of reward delivery to the time of the cue presentation. This response parallels the behaviorally observed conditioned response to the cue, which is predictive for reward delivery. Thus, dopaminergic responses to rewards depend crucially on the rewards’ unpredictability and reflect not reward per se, but reward prediction errors (Schultz et al., 1993, 1997; Schultz, 1998). Similarly, single-unit activity in the orbitofrontal cortex (OFC) has been shown to reflect the reward-predictive properties of stimuli that precede reward delivery (Schoenbaum et al., 1998; Tremblay and Schultz, 1999).

Subsequent fMRI studies in humans found responses to reward and prediction errors in regions to which dopamine neurons project. These studies have identified the striatum and the OFC in the context of primary rewards (Berns et al., 2001; O’Doherty et al., 2002, 2003; McClure et al., 2003) and secondary rewards such as money (Knutson et al., 2001a, 2001b; O’Doherty et al., 2001; Elliott et al., 2003). Further research has shown that signals in the striatum and the OFC comply with principles of formal learning theory and can explain conditioning effects in individual learning such as the blocking phenomenon (Tobler et al.,

2006). More specifically, both regions have been found to show lower responses to blocked as compared with non-blocked, reward-predicting stimuli.

1.4 Neural correlates underlying social learning

In addition to learning from direct experience, an important factor that guides reward-directed behavior is learning from others. Learning by observation is highly functional for an individual as it allows for learning without the costs of direct experience (Boyd et al., 2011; Rendell et al., 2011). Despite the importance of learning by observation the underlying neural mechanisms only recently have become the focus of scientific investigations. Several studies have shown that observing others receiving rewards activates similar regions that have been implicated in processing the direct experience of stimuli and outcomes. For example, one study showed that the degree to which individuals found it rewarding to observe others win monetary rewards was correlated with increased activity in the ventral striatum that overlapped with activations elicited when participants themselves won these rewards (Mobbs et al., 2009). Several further studies also suggested that social learning may be represented in the same brain regions that are relevant for learning through personal experience. One study investigated prediction errors derived from changing the donations for a charity when altruistic choices were made (Kuss et al., 2013). Here, reward prediction errors due to outcome changes for the charity elicited responses in the ventral striatum, similar to the responses elicited by prediction errors due to changes in the outcome for the participants themselves.

Moreover, receiving social rewards during learning in the form of social feedback engages regions, such as the striatum and the ventromedial prefrontal cortex (vmPFC) that are also active during receipt of non-social rewards (Jones et al., 2011; Lin et al., 2012). Similarly, observational outcome prediction errors have been found in the vmPFC (Burke et al., 2010), a region implicated in processing non-social outcome prediction errors (O'Doherty et al., 2003).

However, there are also studies suggesting that social learning involves different neural structures outside the classic reward system. For example Behrens et al. (2008) investigated social learning through advice and found that activations in the dorsomedial prefrontal cortex (dmPFC) and superior temporal sulcus regions within the temporoparietal junction (STS/TPJ) correlated with errors in the predicted trustworthiness of a confederate. Observational action prediction error signals, reflecting the actual minus the predicted choice of others, have been observed in the dorsolateral prefrontal cortex (dlPFC; Burke et al., 2010) and in the dmPFC

during learning through simulation (Suzuki et al., 2012). Another study found that while the medial PFC tracks the predicted reward given the expected influence the participants' choices have on the partner, the posterior STS was responsible for updating the influence signal by the difference between expected and actual influence exerted (Hampton et al., 2008). Thus, there is evidence that at least some forms of social learning engage brain areas not commonly involved in individual learning but which have been implicated in social cognition (Saxe, 2006; Ruff and Fehr, 2014). However, as none of these studies directly compared the social condition with a non-social control condition, the question remains open as to whether, and which, of these regions represent specific involvement in social learning.

While these studies provided useful insights into the neural mechanism that are engaged in social prediction error processing during learning, the representation and implementation of other formal principles during social learning has remained unclear. One important principle is that of efficiency. This optimality principle is illustrated by the blocking effect, which has been studied so far only when individuals learn about their own outcomes but not when they learn about socially relevant outcomes. Moreover, it has yet to be shown whether (and how) this efficiency principle would be abolished when individuals learn in a social context. The study of social learning is not only relevant because of the high functionality of social learning for reward-guided behavior but also because it forms the basis for social decisions that involve higher level cognitive processes. More specifically, social learning can influence moral decisions that affect the outcomes of another individual or of a society as a whole.

1.5 Moral decisions and individual differences

Probably one of the most common moral decisions individuals face when interacting with others is the decision regarding honesty, which is pervasive not only in everyday situations but also in contexts exerting great economic and political impact. How humans think about right and wrong in such moral contexts has received attention in many fields, including those of philosophy, psychology, economics, and law (Goodenough and Prehn, 2004). Two branches of moral philosophy have been most influential in proposing principled answers to this question. Utilitarianism (or consequentialism) states that when determining the moral quality of an action or principle one should consider solely the foreseen consequences of that action or principle. For example, if acting in a certain way maximizes the greater good, that action is morally permissible (Mill, 1861). By contrast, deontological (or non-consequentialist) theories argue that it is more

important that the action should follow certain rules or principles deemed to be morally valuable in themselves. Thus, impermissible actions or principles remain wrong no matter how good their consequences are.

The most influential deontological moral theory has been proposed by Kant, who introduced the idea of the categorical imperative, stating that one should act according to maxims that could become universal laws (Kant, 1965/1785). A concept that captures deontological commitments from a more psychological perspective is that of protected values, which is a shorthand for the finding that at least with some moral decisions some individuals refrain from making trade-offs about consequences because they apply specific moral principles. These principles can be thought of as "protected values" that shield moral decisions against considering economic and other consequences (Baron and Spranca, 1997).

Not surprisingly, moral decisions, such as the ones concerning honesty, vary greatly between humans, for example because individuals differ in the extent to which they consider consequences in their decisions. These differences emerge to a stronger extent when the cost of following the principles (e.g. of telling the truth) is higher. That is, for most individuals the benefits of telling the truth will outweigh the costs as long as these costs are low. However, when the costs increase, they will become too high for some individuals, while others will stand by their moral principles and react less to the costs.

1.6 Neural mechanisms underlying decisions concerning honesty

Recent neuroimaging studies identified a network of brain regions that are involved in making honest or dishonest decisions (for review see Sip et al., 2008; Abe, 2009). In spite of the variety of different experimental paradigms used, such as tasks that involved instructed lying (Spence et al., 2001; Ganis et al., 2003; Langleben et al., 2005; Abe et al., 2006) or spontaneous lying (Baumgartner et al., 2009; Greene and Paxton, 2009; Abe and Greene, 2014), the studies consistently revealed contributions of the dlPFC, the dmPFC, as well as the ventrolateral prefrontal cortex (vlPFC) and inferior frontal gyrus (IFG). Most of the studies found these regions to be activated when individuals engaged in dishonest behavior, as compared with non-deceptive control behavior (Sip et al., 2008; Baumgartner et al., 2009; Abe, 2011). Others observed these regions to be activated when participants voluntarily told the truth, as compared with a control condition in which they had no other choice than telling the truth (Sip et al., 2010), particularly in those participants who lied in most trials (Greene and Paxton, 2009).

These studies looked primarily at the neural activations underlying honesty in group analyses, without directly considering individual differences in the propensity to tell the truth. To date, only one study has explored the neural mechanisms underlying individual differences in spontaneous dishonest behavior. Abe and Greene (2014) found that individuals exhibiting stronger nucleus accumbens responses to anticipated reward show higher levels of dishonest behavior and exhibit greater dlPFC activation when refraining from dishonest behavior. While this study provides important insights in how the lure of economic incentives influence decisions, it is not known how strong moral principles protect against that lure to enable honesty.

There are different views on how the brain processes moral, as compared to non-moral, choices (Greene and Haidt, 2002; Hauser, 2006). It has been argued that moral decision making relies on brain processes similar to those seen during non-moral decisions (Greene and Haidt, 2002; Shenhav and Greene, 2010). This view implies that such judgments are produced by domain-general neural mechanisms that underpin both moral and non-moral choices. In contrast, it has been hypothesized that moral decisions require dedicated neural processes that are specialized for moral behavior (Hauser, 2006; Mikhail, 2007). Accordingly, these processes are thought to be distinct from those involved in non-moral decisions and therefore represented by a domain-specific moral faculty. It is an open question how individual differences in behavior and protected values might relate to a neural activity pattern that is specific to moral as compared to non-moral decisions.

2 Overview of the studies

2.1 Study 1: Neural correlates of efficient learning mechanisms in the social domain

Background

Recent work has extended formal models of individual learning to the social domain and investigated social prediction error processing during social learning (Behrens et al. 2008; Hampton et al., 2008; Burke et al. 2010; Suzuki et al. 2012). However, the representation and implementation of other formal principles during learning in social situations has remained unclear. One principle important for optimal learning is that of efficiency and is represented by the blocking effect. Blocking of learning occurs when a novel cue is presented together with a previously learned cue that predictably leads to the same outcome as the novel cue. Learning about the novel cue is blocked because the previously learned cue has already been established as a reliable predictor of the outcome. Thus learning about the novel cue would be inefficient and redundant. The blocking effect is expressed in a diminished behavioral response to novel stimuli that provide only redundant information about reward occurrence, as compared with novel stimuli that provide non-redundant information. While blocking is established for individual learning, it is not clear how this efficiency principle extends to reward learning in the social domain.

Cues that are redundant for others might nevertheless be of value to us as they may provide us with relevant information that is either relevant now or will become relevant later. In particular, these cues may enter different associations with rewards for ourselves, as compared with rewards for others. Therefore, we might want to keep track of social cues, even if they are redundant for others. By extension, whether we still rely on previous learning (resulting in blocking) or track all information available to us in the environment (not resulting in blocking) remains an open question for reward learning in the social domain.

If the blocking effect indeed occurred in the social domain, the question arises about whether its implementation depends on the same regions that implement it in the individual domain. Alternatively, specific regions may be involved in blocking redundant reward learning in the social domain that nevertheless implement the blocking effect by following similar computational principles. By investigating these possibilities we may be able to illuminate whether the neural representations of social learning are unique and different from those of

individual learning. To address this question we employed a social variant of the blocking paradigm together with fMRI. In particular, we examined the specific role of the mPFC in efficient learning about social monetary rewards by investigating individual differences in social blocking.

Methods

With thirty-eight healthy participants we investigated the blocking effect by using social and individual monetary rewards in two separate conditions. Individual rewards were received by the participant, social rewards by another person. We used a within-subject version of the blocking procedure (Waelti et al., 2001; Tobler et al., 2006) that comprised three consecutive phases. In each of these phases, participants were presented with visual stimuli that were associated with the delivery of different social or individual outcomes. In the first (pretraining) phase, the A (experimental) stimuli (A_{SOCIAL} and $A_{INDIVIDUAL}$) were paired with a social or individual reward. In contrast, the B (control) stimuli (B_{SOCIAL} and $B_{INDIVIDUAL}$) were not paired with a reward. During the presentation of any given stimulus, the participants were to perform a specific key press corresponding to the recipient and to the outcome that would follow the stimulus. fMRI scanning started in the second (compound conditioning) phase. A stimuli were presented together with X stimuli (X_{SOCIAL} and $X_{INDIVIDUAL}$), forming rewarded compounds. As a control, B stimuli were presented together with Y stimuli (Y_{SOCIAL} and $Y_{INDIVIDUAL}$) and also followed by a reward. In a third phase, X and Y stimuli were presented alone in unrewarded test trials. Under the assumption that previous learning blocks subsequent learning, the X stimuli should have been blocked from being associated with social or individual reward, while the Y stimuli should have been associated with reward. Social and individual reward expectation was defined as the percentage of the social and individual reward key pressed, respectively. The degree of participant-specific behavioral blocking was calculated as the difference between recipient-specific reward key presses for Y stimuli and those for X stimuli. The larger the difference, the stronger the blocking effect.

fMRI data processing and statistical analyses were carried out using statistical parametric mapping (SPM8). Data preprocessing consisted of realignment, coregistration, segmentation, spatial normalization, and smoothing. Data analysis was performed using a general linear model (GLM) approach. The first-level design matrix of each participant included separate regressors for each of the four learned stimulus conditions ($A_{INDIVIDUAL}$, A_{SOCIAL}, $B_{INDIVIDUAL}$, B_{SOCIAL}) modeled at the event-onset time, the compound conditioning trials at the time of the outcome

($AX_{INDIVIDUAL}$, AX_{SOCIAL}, $BY_{INDIVIDUAL}$, BY_{SOCIAL}) to capture prediction error-related responses during learning, and the four test trial types modeled at the event-onset time ($X_{INDIVIDUAL}$, X_{SOCIAL}, $Y_{INDIVIDUAL}$, Y_{SOCIAL}) to capture blocking. We parametrically modulated the AX and BY regressors with trial-wise and mean-corrected prediction errors derived from a standard reinforcement learning model. Linear contrasts of regression coefficients of A vs. B (stimulus response), BY vs. AX (prediction error modulator), and Y vs. X (stimulus response) were computed at the single-participant level and then taken to group-level analyses where we used one-sample t-tests or correlations with participant-specific degree of blocking in the social or individual domain.

Results & Conclusions

We observed learning and blocking not only in the individual but also in the social condition as evidenced by a significantly higher number of reward key presses for A and Y stimuli as compared to B and X stimuli. Moreover, for both conditions we found individual differences in the degree of blocking and these differences were correlated between the individual and social condition. Despite this similarity on the behavioral level we found a preferential role for the relatively more dorsal part of the PFC for social learning and blocking: we found that activity in the mPFC was stronger for reward-predicting stimuli than for neutral stimuli (A vs. B) and that this activity increased with the degree of behavioral social blocking (Y vs. X). Thus, within mPFC, similar subregions showed activations reflecting reward expectation and blocking in the social domain. Importantly, these activations in the mPFC for social learning and blocking were enhanced over and above those in the individual domain (which were represented in ventral regions of the mPFC).

Additionally, we found that in the BY conditions, reward key presses gradually increased, but that there was only a very small increase in reward key presses in the AX conditions as the reward outcome for AX was already fully predicted by the pretrained A stimuli. Consequently, the reward outcome was expected to generate a sizeable prediction error in early BY trials. As learning progressed from trial to trial, the reward outcome was expected to elicit a gradually decreasing prediction error in later BY trials. Activation in the dmPFC fitted the parametric modulator for BY compared with AX trials, better. This reflects the stronger decreasing prediction error responses in BY trials, specifically in the social condition.

Taken together, the findings suggest that the efficiency principle that applies to reward learning in the individual domain also applies to that in the social domain, with the mPFC playing a central role in implementing it.

2.2 Study 2: Social unblocking

Background

Previous studies have shown that blocking can be prevented if the value (but not the identity) or the identity (but not the value) of the outcomes is changed (Burke et al., 2008; McDannald et al., 2011). In contrast to blocking, changing the value or identity of the outcome results in learning of the second cue due to the discrepancy between the expected and the actual value or identity of the outcome. However, several studies (Bakal et al., 1974; Dickinson and Dearing, 1979; Ganesan and Pearce, 1988) have shown that changing the identity but not the value of the outcome can result in a similar degree of blocking. This is the so called "transreinforcer blocking" phenomenon.

Transreinforcer blocking and unblocking have been studied in rats (Burke et al., 2008; McDannald et al., 2011; Steinberg et al., 2013) and in humans during causal learning (Le Pelley et al., 2005) when learning took place with rewards received by the learning individual. However, in real life we often learn in an environment in which others receive rewards as well. It is not clear whether or not social outcome shifts would block learning. More specifically, it is an open question how learning would take place if neither the subjective value, nor the identity of the outcome change, but its recipient. Changing the recipient of a reward might result in unblocking, or alternatively, as the value and identity of the outcome are the same, lead to transrecipient blocking. To address this question, we developed a transrecipient variant of the (un)blocking paradigm in which we changed the recipient of the monetary rewards while keeping their subjective value unchanged across individual and social conditions.

How we learn about social information might depend also on factors that are not related to the outcome itself, but to how we feel towards others and to the degree to which we experience similar feelings from social and individual outcomes. Thus, social learning might depend on our capacity to experience empathy, a relationship that has been demonstrated in recent studies (Fukushima and Hiraki, 2009; Rak et al., 2013; Gossen et al., 2014). We therefore assessed participants' empathy and hypothesized that people high in empathy will be less sensitive to

recipient changes, as they perceive the consequences of others' outcomes similarly to the consequences of their own outcomes.

Methods

Thirty-four participants performed a blocking (see Study 1) and a transrecipient (un)blocking experiment in two separate sessions, on different days. Like the blocking experiment, the transrecipient (un)blocking experiment comprised three phases. The first, pretraining phase in the (un)blocking experiment was identical to the pretraining phase in the blocking experiment. However, the second, compound conditioning phase differed from the standard blocking experiment. Here, the reward recipient was changed when the pretrained stimuli A were presented in compound with stimuli X. In the social (un)blocking condition a $A_{INDIVIDUAL}$ stimulus was pretrained with a monetary outcome for the participant, but then resulted in a reward (of equal value) for another person when presented with a second stimulus during compound conditioning (AX_{SOCIAL}). Conversely, in the individual (un)blocking condition pretrained rewards were received by another person (A_{SOCIAL}) and the recipient changed to the participant in the compound phase ($AX_{INDIVIDUAL}$). To keep the (un)blocking experiment as similar as possible to the blocking experiment, there were also rewarded BY control conditions, although they were not analyzed further. In these conditions, after participants had learned that control stimulus B predicted no reward for one recipient (e.g. individual), the presentation of stimuli B and Y in compound was followed by reward for the other recipient (e.g. social). As in the blocking experiment, in the third phase, X and Y stimuli were presented alone in unrewarded test trials. Thus, other than the change of reward recipient in the compound conditioning phase, the transrecipient (un)blocking experiment was identical to the blocking experiment with respect to trial structure and trial number. After completion of the experiment, we assessed participants' trait empathy (Interpersonal Reactivity Index (IRI); Davis, 1983) and prosociality (Caprara et al., 2005).

Results & Conclusions

When including all subjects, irrespective of their behavior in the blocking experiment, we observed that stimulus X was unblocked from learning when an individual reward outcome changed to a social reward outcome, as participants showed increased reward-expecting button presses for stimulus X in the (un)blocking experiment as compared with the blocked stimulus X in the standard blocking experiment. By contrast, when a social reward outcome changed to

individual reward our data indicate a transrecipient blocking effect as responses to stimulus X in the unblocking experiment do not differ significantly from the responses to stimulus X in the blocking experiment.

When only investigating participants who showed blocking in the blocking experiment, we observed unblocking effects for both the individual and social conditions and these unblocking effects were correlated between conditions. Moreover, we found empathic concern and prosociality to be negatively correlated with the degree of unblocking in both the individual and social condition, indicating that participants with less empathic concern and prosociality showed more unblocking. Taken together, our findings suggest that value-matched learning is sensitive to the reward recipient and that individual differences in unblocking may relate to social traits.

2.3 Study 3: Brain connectivity and individual differences in honesty

Background

Decisions concerning honesty are one of the most common moral decisions. As with many other moral decisions, humans show profound individual differences in their decisions about honesty. These differences are reflected for example in personality traits that determine how individuals will behave when telling the truth is costly (Gibson et al., 2013). Protected values (Baron and Spranca, 1997) represent such a personality trait. They can be thought of as principles that shield moral decisions against consideration of economic and other consequences. Recent neuroimaging studies have identified a network of brain regions in the PFC that are involved in decisions concerning honesty (Sip et al., 2008; Abe, 2009). However, it is not clear how these brain regions dynamically communicate with each other in response to variations in the cost of telling the truth and whether these communications depend on individual differences in protected values.

To address these questions we used fMRI to measure brain activations while participants made decisions concerning truthfulness. It is likely that complex behavioral processes, which depend on personality trait and situational costs, are not reflected by activity of a single brain region, but by dynamic changes in the connectivity among different brain regions. Specifically, we hypothesized that individuals with high protected values may exhibit a specific connectivity pattern that protects them against the lure of economic incentives when the costs of telling the truth are high. Our experimental design also allowed us to address a central question in the study

of morality, namely, the extent to which moral and non-moral decision making relies on similar versus different neural mechanisms (Greene and Haidt, 2002; Hauser, 2006). We hypothesized that particularly individuals high in protected values tap into a specific domain.

Methods

Thirty-two participants were placed in the situation of an imaginary chief executive officer (CEO) who had to announce the company's earnings. This setup parallels a real-life conflict that CEOs face, as their variable compensation is often tied to stock price performance, which, in turn, depends on the earnings they announce. Thus, despite knowing that it is ethically wrong, the CEO has economic incentives to behave dishonestly and boost earnings in order to increase his or her payout. Importantly, the CEOs' preferences for telling the truth should decrease as the costs of telling the truth increase.

Our experimental design aimed to create a comparable situation by varying the economic cost of truthfulness, which in turn allowed participants to trade off economic with moral incentives. Thus, in each trial of the moral task, participants had to choose whether to honestly announce true earnings or to lie and falsely report higher earnings. Importantly, false reports led to a high, fixed payoff, whereas truthful reports led to lower actual payoffs, which were varied parametrically from trial to trial. Moreover, we included control tasks with comparable trade-off situations that did not contain a moral aspect. In the valuation task participants chose between two projects to invest in, where one of the options yielded a higher profit than the other one. In the effort task, participants had to choose how much work to invest as a CEO, such that one option yielded a higher profit but required more work than the other option.

In each trial of each task, participants first saw the variable, economically less beneficial option followed by the constant, economically more beneficial but false or more effortful option. The second, constant option was presented together with the first option during which participants had to indicate their choice. The cost of choosing the first option varied between 0 and 4 CHF. Both task type and payoff levels were randomly intermixed across trials during the experiment. After scanning, we assessed to what extent participants treated truthfulness as a protected value and felt committed to telling the truth with a questionnaire consisting of two subscales (direct and indirect protected values) as described previously (Tanner et al., 2009; Gibson et al., 2013).

fMRI data processing and statistical analyses were carried out using SPM8. Data preprocessing consisted of realignment, coregistration, segmentation, spatial normalization and

smoothing. The GLM identified brain regions in which activity correlated with cost-level by using the following set of regressors: presentation of first Truthtelling option, presentation of first Valuation option, and presentation of first Effort option. For each of the regressors we included a parametric modulator capturing the variable cost-level. Main effects for cost were computed on the single-subject level by performing separate t-tests for each parametric modulator. The resulting contrast images were taken up to the group-level where we used correlations with participant-specific percentage of truthtelling (or low value project/effort chosen) and protected values. Further, we used psycho-physiological interaction (PPI) analyses (Friston et al., 1997) to test for coupling differences due to variations in protected values. For this analysis we adapted the first GLM and computed a first-level model in which we defined separate regressors depending on cost. With the average time course extracted from voxels in the dlPFC and dmPFC as physiological regressor, we performed a first-level model with the cost-level as a psychological regressor (high vs. low cost), and a psycho-physiological interaction (PPI) regressor. We then used the PPI regressor to perform a second-level correlation analysis with individual protected values.

Results & Conclusions

Participants showed a wide variation in their choices, with some participants choosing to tell the truth in all trials and others doing so in very few trials. We observed an interaction between the strength of participants' direct protected values and economic costs of truthfulness suggesting that, when economic costs of truthfulness were high, participants with stronger protected values chose the truthful option more often, compared with participants with weaker protected values. In the two control tasks we found no interaction effect that followed such a pattern, indicating that high protected values reduced participants' willingness to trade off specifically moral values rather than non-moral values against economic costs.

On the neural level we found that coding the cost of honesty increased with the individual degree of honesty in the dlPFC and dmPFC. In other words, with increasing cost of telling the truth participants showed stronger parametric cost-related activation in the dlPFC and dmPFC the more often they made honest decisions. When testing for specificity we observed no significant activation specific for the honesty task. This finding is in line with the domain-general notion that cost coding in the dlPFC and dmPFC is related to the individual propensity of incurring any kind of costs, rather than specifically to the propensity of incurring economic costs for moral benefits.

As our behavioral analyses indicated that higher protected values support honest decisions when the cost of telling the truth increases, we investigated whether protected values influence the interaction of other regions with the dorsolateral and dorsomedial prefrontal regions that code costs as a function of individual honesty. We found that functional connectivity between the dlPFC and the inferior frontal gyrus (IFG), as well as between the dmPFC and the IFG, differed significantly as a function of cost-level and protected values. More specifically, both regions showed stronger functional connectivity with the IFG in high cost conditions, compared with low-cost conditions, as protected values increased. Moreover, cost-level dependent connectivity between dlPFC and IFG related more strongly to protected values in the honesty task than in the other two tasks. Thus, protected values appear to exert their effects on moral decision making particularly via a connection between the dlPFC and the IFG. Taken together, our results provide the first evidence that neural connectivity patterns during decisions involving honesty are modulated by individual differences in protected values.

3 General Discussion

3.1 Study 1

The first study revealed that the efficiency principle represented by the blocking effect extends to reward learning in the social domain. Importantly, the more dorsal mPFC implements this effect specifically in social contexts. Thus, this region contributes to other-directed reward learning through an efficient learning mechanism originally described in empirical studies and formal models of individual learning (Kamin, 1969; Rescorla and Wagner, 1972). Accordingly, social and individual learning seem to engage neural processes that follow similar computational principles but are implemented in distinct areas of the brain.

The presented results converge with reports of relatively dorsal mPFC involvement in other aspects of social learning. For example, Behrens et al. (2008) found that activity in the dmPFC correlates with errors in the predicted credibility of confederate advice. In another study, dmPFC activation during an inspection game was found to correlate with the degree to which players thought their own action influenced their opponent's chosen strategy (Hampton et al., 2008). Thus, activity of the dorsal mPFC can be captured particularly well with formal models of social learning with the unifying explanation that this region encodes social reward prediction errors. This is in line with the findings of previous studies which showed that the dorsal mPFC incorporates factors that are relevant for learning about stimuli and outcomes in social situations. Activity in this region has been found to be related to social value processing, other-related judgments, and inferring the mental states of others (Ochsner et al., 2004; Amodio and Frith, 2006; Gilbert et al., 2006; Mitchell, 2009; Krienen et al., 2010; Fareri et al., 2012). All of these processes come into play when we learn in a social context. More specifically, social value processing, mentalizing and thinking about others constitute the motivational and cognitive factors that are relevant for learning about stimuli and outcomes that concern others.

Outcomes received by others may be more abstract than one's own outcomes (Amodio and Frith, 2006). In this sense, the present findings support the idea of a dorsal-ventral and posterior-anterior axis (Denny et al., 2012; Suzuki et al., 2012; Koritzky et al., 2013) according to which the more dorsal and anterior mPFC processes more abstract and complex information than the more ventral and posterior mPFC. From an evolutionary point of view, the results are consistent with the notion that the anterior part of the prefrontal cortex may have emerged as a new prefrontal region during primate evolution (Genovesio et al., 2014). Accordingly, the

complexity of social environments increased and social functions developed to a disproportionate degree in the later stages of primate evolution (Dunbar, 1998). It might be the case that these social aspects of the environment are processed in brain regions that evolved specifically to deal with these demands. Thus, it is tempting to speculate that the very frontal part of the mPFC might have evolved to serve a preferential role for learning about observed and socially relevant outcomes. In line with this, an interesting avenue for future research would be to elucidate how the efficiency principle during social learning is represented in different non-human primates and whether it can be mapped onto the development of the anterior mPFC.

There have also been reports of other-relevant learning processes in the vmPFC (Burke et al., 2010; Suzuki et al., 2012). One possibility worthy of further study is that these ventral regions are engaged when other-relevant learning has a direct benefit (instrumental value) for the observing individual (Burke et al., 2010). In contrast to that study, the observation of others' rewards had comparatively little instrumental value in our study. Thus, it remains to be determined what specific aspects determine whether the vmPFC contributes to learning in the social domain. Personal relevance or instrumental value appear to be promising candidates.

The dmPFC activation that reflected the gradual decrease in social prediction errors in BY trials was more dorsal and posterior than the mPFC region, which was sensitive to social blocking. This suggests that different subregions of the dmPFC are engaged at different stages of social learning. Future research may therefore focus on the mechanisms underlying the development of blocking in the social domain and investigate in more detail how the development of the effect in the compound phase relates to its expression in the test phase.

To conclude, the present results advance our knowledge about social learning and demonstrate the informative role of neuroscientific approaches, particularly in cases in which behavior is similar for social and individual learning, but the neural mechanisms underlying these two types of learning are different. By comparing social with individual learning, we can begin to dissociate them at the neural level and assess whether social learning is unique in terms of behavioral and computational mechanisms.

3.2 Study 2

The second study investigated the unblocking effect in a social context by using a novel paradigm in which the recipient of the monetary outcome changed without accompanying changes in subjective value or reward identity (money). We found that participants who normally

block redundant learning in the standard blocking paradigm show unblocking when the reward recipient changes. Thus, although social and individual learning follow similar learning principles and are associated with correlated blocking effects, humans remain sensitive to who receives reward, and they track value-preserving recipient changes by showing unblocking.

When investigating the behavior of all participants, we observed unblocking for the social but not the individual condition, suggesting that unblocking occurs preferentially when individual rewards change to socially relevant rewards. This finding seems to contradict previous work showing a bias toward a reliance on individual information, at least when individual information competes with social information (Eriksson and Strimling, 2009; Morgan et al., 2012). A bias favoring individual information would predict that participants should rely more on individual cues and outcomes when the outcome changes from individual to social. Thus, if they were more sensitive to learning from individual information participants should have shown unblocking when reward receipt changed from social to individual and blocking when receipt changed in the reverse direction. It is conceivable that the degree of reliance on individual information depends on the degree of competition between individual and social information but future research is needed to elucidate this possibility.

Interestingly, lower scores in empathy and prosociality were associated with greater unblocking effects, suggesting that less empathic and prosocial individuals are more sensitive to who actually receives rewards. In other words, more empathic concern leads individuals to treat others' rewards more similarly to their own and higher selfishness leads to increased differentiation between rewards received by self and other. This interpretation is supported by previous research that has linked empathy to altruism (Batson et al., 1991, 2003) and sharing behaviors (Edele et al., 2013). Our empathy finding should be interpreted with caution, however, since we only find a correlation with the empathic concern subscales of the IRI. In particular, our initial hypothesis would predict that also the perspective taking subscale should relate to unblocking effects. Empathy is currently thought of as a multidimensional concept, involving a cognitive and an affective component (Davis, 1983; Decety and Lamm, 2006). While perspective-taking represents the cognitive aspect, empathic concern reflects the affective aspect of empathy. Thus, our data seem to indicate that not understanding others' perspective but the emotional reaction to others' needs and welfare accounts for differences in the degree of unblocking in a social context. The findings of the second study might also have implications that go beyond social learning. Specifically, whether the recipient of outcomes is exchangeable during learning might influence any kind of decision with social outcomes. For example,

individuals might more strongly follow moral principles that increase others' welfare because they learn about others' outcomes as if they were their own. Thus, individual differences in social decision making might depend largely on how we learn about social outcomes in the first place, which in turn may reflect how empathic we are.

3.3 Study 3

In the third study we investigated how individual differences in behavior and protected values are represented in the brain, particularly during moral decisions involving honesty. Specifically, we revealed how functional coupling between prefrontal regions changes as a function of honesty costs and protected values. We found that in high, compared with low-cost situations, the dlPFC and dmPFC show stronger coupling with the IFG with increasing levels of protected values.

Given that these regions have been consistently implicated in cognitive control and response inhibition (Botvinick et al., 2001; Miller and Cohen, 2001; Aron, 2007; Carter and van Veen, 2007), the increased dlPFC and dmPFC responses, and their connectivity with the IFG might represent an active control mechanism that helps honest individuals to refrain from dishonest behavior, specifically when economic costs increase. These regions have also been found to play a crucial role during decisions involving honesty. While some studies showed that the dlPFC (Greene and Paxton, 2009; Abe and Greene, 2014) and dmPFC (Greene and Paxton, 2009) are reflecting honest behavior, the majority of studies reported these regions to be associated primarily with deceptive behavior (Sip et al., 2008; Abe, 2009). This primary involvement in deceptive behavior has been taken as evidence that lying is cognitively more demanding and thus requires more cognitive control than telling the truth. Our results reconcile these discrepant findings by showing that the engagement of control regions during honesty decisions depends on both the decision situation and on individual differences between decision makers. Interestingly, recent behavioral research has shown that active self-control does indeed support honest behavior (Mead et al., 2009; Gino et al., 2011; Shalvi et al., 2012). Taken together, our data suggest that stronger protected values could be linked to stronger self-control mechanisms.

Our results reveal that the connectivity in individuals with strong protected values is enhanced specifically when high economic costs are involved and thus strong control mechanisms are needed to protect the moral values. It has been proposed that protected values

derive from deontological decision rules that require or prohibit certain actions (Baron and Spranca, 1997). On a very basic level these rules might be stored as semantic knowledge, which is retrieved when individuals are faced with decision situations that are associated with those rules. Interestingly, the IFG has been implicated in semantic rule retrieval and processing (Bunge, 2004; Badre et al., 2005; Souza et al., 2009). Thus, it might be that the input to the dlPFC and dmPFC provided by the IFG represents semantic rules that help enforce honest behavior (although it should be kept in mind that directionality cannot be inferred from PPI effects). In agreement with this interpretation, a recent study has shown responses in the IFG when individuals refused to give up moral values in exchange for economic benefits (Berns et al., 2012).

Our results pertain to the central question of whether moral decisions are represented by domain-specific or domain-general neural mechanisms (Greene and Haidt, 2002; Hauser, 2006). Our findings suggest that the correlation between cost coding in the dlPFC and dmPFC with individual honesty is not specific to moral trade-off situations, but represents a more general relation of coding the costs of a decision to individual differences in decisions that involve any kind of perceived trade-offs. By contrast, individual differences in protected values predict the strength of dlPFC-IFG connectivity specifically in the honesty task, but not in the two control tasks. Thus, differences that arise from individually varying protected values specifically relate to connectivity patterns during moral as opposed to non-moral decision making. These findings imply that the brain regions involved in moral decision making are not specific to moral decisions, but the way in which they interact with each other, as a function of cost and protected values, is specifically related to moral aspects of decisions.

Previous studies mostly focused on the question of *where* in the brain honest decisions, or more generally, moral decisions (Young and Dungan, 2012) are processed. Our results indicate that a complete understanding of decisions involving honesty requires scientists interested in the neuroscience of moral behavior not only to identify single activated brain regions, but also to determine *how* brain regions dynamically interact with respect to personality traits and economic costs of the moral behavior. Advanced analysis methods, such as dynamic causal modeling, may help determine directionality of the present connectivity patterns. Moreover, causal methods, such as alternating current stimulation, could be used to explore whether the observed connectivity pattern can be changed and whether this leads to changes in the inter-individual variability of honesty decisions.

Taken together, our results provide the first evidence that neural connectivity patterns during decisions involving honesty are modulated by individual differences in protected values. Individual variability in behavior modulates cost coding in the dlPFC and dmPFC, both of which were similarly involved during more general, non-moral decisions. Individual differences in protected values modulate the dlPFC-IFG connectivity specifically during moral decision making. Our results thereby highlight the role of individual differences in moral attitudes and behavior, and provide an explanation as to why some people decide to follow a moral principle whereas others do not.

3.4 General Conclusions

Many current models of brain function are built on commonalities across individuals and unique activity patterns are often considered as noise (Thompson-Schill et al., 2005; Van Horn et al., 2008). The studies presented in this thesis show how the analysis of individual differences can provide useful insights into the behavioral and neural mechanisms of social learning and moral decision making. Our findings have several notable implications. First, the results demonstrate that individual differences exist and are large. For example, study 1 showed great variability in the degree of blocking. Similarly, in study 3 some subjects almost always lied, while others always told the truth. Second, individual differences pervade mental functions. This was demonstrated for instance in study 1, where we found correlated blocking effects for individual and social learning. Third, the investigation of individual trait differences can reveal psychological factors that lead to individual differences in behavior. For instance, in study 2 empathy scores predicted the degree of unblocking, suggesting that the degree of social transrecipient blocking or unblocking depends on the ability to share feelings experienced by self and other. Fourth, individual differences in social behavior are consistently expressed in the prefrontal cortex. This was illustrated in study 3, where prefrontal connectivity predicted individual differences in moral disposition and in study 1, which revealed that individual differences in social blocking corresponded with differences in mPFC activation. Given that the prefrontal cortex is connected to most cortical and subcortical structures and targets main sources of neurotransmitter systems in the basal forebrain and brainstem (Goldman-Rakic, 1987; Pandya and Yeterian, 1990), it seems that it is particularly well suited to express individual differences in social behavior. Together, our findings demonstrate that the study of individual differences leads to a more comprehensive understanding of the behavioral and neural underpinnings in

social learning and moral decision making and imply that taking variability into account allows more precise predictions about social behavior.

The study of individual differences in the domain of social learning and moral decision making is also relevant for improving our understanding of various psychiatric disorders that involve abnormal social behavior, such as schizophrenia, autism, and psychopathy. As many of the psychiatric diseases are on a continuum with variability in the healthy population we could understand disorders, in part, through understanding normal individual differences. For example, schizophrenic patients show reduced blocking, suggesting that prediction error processing is different in these patients (Jones et al., 1997; Fletcher and Frith, 2009). However, so far, studies have focused on altered blocking mechanisms in the context of individual learning (Oades et al., 1996; Jones et al., 1997; Bender et al., 2001). Given that schizophrenia patients show various deficits in the social domain (Couture et al., 2006; Penn et al., 2008) it could be illuminating to investigate the blocking effect in these patients also with respect to social outcomes. This could be helpful for tracing impairments in social functioning and relating the deficits to specific stages of the disorder. Further, there might be differences between the various psychiatric disorders that involve social deficits. For example, it can be hypothesized that schizophrenia patients do show diminished blocking in both the individual and social domain, but autistic patients exhibit reduced blocking only in the social domain.

The findings of the third chapter also make suggestions for the study of extreme cases of immoral behavior, as seen in psychopathy and antisocial personality disorder. Our findings point out the relevance of studying prefrontal connectivity, as individual differences in prefrontal coupling might reflect patterns that are responsible for predisposing some individuals to pathological moral behavior. Specifically, psychopaths might be more sensitive to costs and exhibit weak protected values, which might explain the high levels of deceptive behavior observed in psychopaths (Hare, 1998). On the neural level, the present results would suggest that psychopaths, as compared with healthy controls show reduced connectivity with the IFG when the costs of telling the truth are high.

In conclusion, our findings imply that the study of individual differences in social learning and moral decision making can inform other fields that investigate abnormal social behavior and suggest that for some research questions inter-individual variability should be treated as data, and not as noise (Kosslyn et al., 2002; Thompson-Schill et al., 2005).

References

Abe N (2009) The neurobiology of deception: evidence from neuroimaging and loss-of-function studies. Curr Opin Neurol 22:594–600.

Abe N (2011) How the brain shapes deception: an integrated review of the literature. Neuroscientist 17:560–574.

Abe N, Greene JD (2014) Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. J Neurosci 34:10564–10572.

Abe N, Suzuki M, Tsukiura T, Mori E, Yamaguchi K, Itoh M, Fujii T (2006) Dissociable roles of prefrontal and anterior cingulate cortices in deception. Cereb Cortex 16:192–199.

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268–277.

Aron AR (2007) The neural basis of inhibition in cognitive control. Neuroscientist 13:214–228.

Badre D, Poldrack RA, Paré-Blagoev EJ, Insler RZ, Wagner AD (2005) Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. Neuron 47:907–918.

Bakal CW, Johnson RD, Rescorla RA (1974) The effect of change in US quality on the blocking effect. Pav J Biol Sci 9:97–103.

Baron J, Spranca M (1997) Protected values. Organ Behav Hum Dec 70:1–16.

Batson CD, Batson JG, Slingsby JK, Harrell KL, Peekna HM, Todd RM (1991) Empathic joy and the empathy-altruism hypothesis. J Pers Soc Psychol 61:413–426.

Batson CD, Lishner DA, Carpenter A, Dulin L, Harjusola-Webb S, Stocks EL, Gale S, Hassan O, Sampat B (2003) "... As you would have them do unto you": Does imagining yourself in the other's place stimulate moral action? Pers Soc Psychol Bull 29:1190–1201.

Baumgartner T, Fischbacher U, Feierabend A, Lutz K, Fehr E (2009) The neural circuitry of a broken promise. Neuron 64:756–770.

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS (2008) Associative learning of social value. Nature 456:245–249.

Bender S, Müller B, Oades RD, Sartory G (2001) Conditioned blocking and schizophrenia: a replication and study of the role of symptoms, age, onset-age of psychosis and illness-duration. Schizophr Res 49:157–170.

Berns GS, Bell E, Capra CM, Prietula MJ, Moore S, Anderson B, Ginges J, Atran S (2012) The price of your soul: neural evidence for the non-utilitarian representation of sacred values. Philos Trans R Soc Lond B Biol Sci 367:754–762.

Berns GS, McClure SM, Pagnoni G, Montague PR (2001) Predictability modulates human brain response to reward. J Neurosci 21:2793–2798.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624–652.

Boyd R, Richerson PJ, Henrich J (2011) The cultural niche: Why social learning is essential for human adaptation. Proc Natl Acad Sci USA 108:10918–10925.

Bunge SA (2004) How we use rules to select actions: a review of evidence from cognitive neuroscience. Cogn Affect Behav Neurosci 4:564–579.

Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) Neural mechanisms of observational learning. Proc Natl Acad Sci USA 107:14431–14436.

Burke KA, Franz TM, Miller DN, Schoenbaum G (2008) The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. Nature 454:340–344.

Bush RR, Mosteller F (1951) A mathematical model for simple learning. Psychol Rev 58:313–323.

Byrom NC (2013) Accounting for individual differences in human associative learning. Front Psychol 4:588.

Caprara GV, Steca P, Zelli A, Capanna C (2005) A new scale for measuring adults' prosocialness. Eur J Psychol Assess 21:77–89.

Carter CS, van Veen V (2007) Anterior cingulate cortex and conflict detection: an update of theory and data. Cogn Affect Behav Neurosci 7:367–379.

Couture SM, Penn DL, Roberts DL (2006) The functional significance of social cognition in schizophrenia: a review. Schizophr Bull 32 Suppl 1:S44–S63.

Davis MH (1983) Measuring individual differences in empathy: Evidence for a multidimensional approach. J Pers Soc Psychol 44:113–126.

Decety J, Lamm C (2006) Human empathy through the lens of social neuroscience. Sci World J 6:1146–1163.

Denny BT, Kober H, Wager TD, Ochsner KN (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24:1742–1752.

Dickinson A, Dearing MF (1979) Appetitive-aversive interactions and inhibitory processes. In: Mechanism of learning and motivation (Dickinson A, Boakes RA, eds), pp 203–231. Hillsdale, NJ: Erlbaum.

Dunbar RIM (1998) The social brain hypothesis. Evol Anthropol 6:178–190.

Edele A, Dziobek I, Keller M (2013) Explaining altruistic sharing in the dictator game: The role of affective empathy, cognitive empathy, and justice sensitivity. Learn Individ Differ 24:96–102.

Eippert F, Gamer M, Büchel C (2012) Neurobiological mechanisms underlying the blocking effect in aversive learning. J Neurosci 32:13164–13176.

Elliott R, Newman JL, Longe OA, Deakin JFW (2003) Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study. J Neurosci 23:303–307.

Eriksson K, Strimling P (2009) Biases for acquiring information individually rather than socially. J Evol Psychol 7:309–329.

Fareri DS, Niznikiewicz MA, Lee VK, Delgado MR (2012) Social network modulation of reward-related signals. J Neurosci 32:9045–9052.

Fletcher PC, Frith CD (2009) Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci 10:48–58.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

Fukushima H, Hiraki K (2009) Whose loss is it? Human electrophysiological correlates of non-self reward processing. Soc Neurosci 4:261–275.

Ganesan R, Pearce JM (1988) Effect of changing the unconditioned stimulus on appetitive blocking. J Exp Psychol Anim Behav Process 14:280–291.

Ganis G, Kosslyn SM, Stose S, Thompson WL, Yurgelun-Todd DA (2003) Neural correlates of different types of deception: an fMRI investigation. Cereb Cortex 13:830–836.

Genovesio A, Wise SP, Passingham RE (2014) Prefrontal-parietal function: from foraging to foresight. Trends Cogn Sci 18:72–81.

Gibson R, Tanner C, Wagner A (2013) Preferences for truthfulness: heterogeneity among and within individuals. Am Econ Rev 103:532–548.

Gilbert SJ, Spengler S, Simons JS, Steele JD, Lawrie SM, Frith CD, Burgess PW (2006) Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. J Cogn Neurosci 18:932–948.

Gino F, Schweitzer ME, Mead NL, Ariely D (2011) Unable to resist temptation: How self-control depletion promotes unethical behavior. Organ Behav Hum Dec 115:191–203.

Goldman-Rakic PS (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: Handbook of Physiology, The Nervous System, Higher Functions of the Brain (Plum F, ed), pp 373–417. Bethesda, MD: American Physiological Society.

Goodenough OR, Prehn K (2004) A neuroscientific approach to normative judgment in law and justice. Philos Trans R Soc Lond B Biol Sci 359:1709–1726.

Gossen A, Groppe SE, Winkler L, Kohls G, Herrington J, Schultz RT, Grunder G, Spreckelmeyer KN (2014) Neural evidence for an association between social proficiency and sensitivity to social reward. Soc Cogn Affect Neurosci 9:661–670.

Greene JD, Paxton JM (2009) Patterns of neural activity associated with honest and dishonest moral decisions. Proc Natl Acad Sci USA 106:12506–12511.

Greene J, Haidt J (2002) How (and where) does moral judgment work? Trends Cogn Sci 6:517–523.

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci USA 105:6741–6746.

Hare RD (1998) Psychopathy, affect and behavior. In: Psychopathy: Theory, Research and Implications for Society (Cooke DJ, Forth AE, Hare RD, eds), pp 105–137. Springer Netherlands.

Hauser MD (2006) The liver and the moral organ. Soc Cogn Affect Neurosci 1:214–220.

Holland PC (1984) Unblocking in Pavlovian appetitive conditioning. J Exp Psychol Anim Behav Process 10:476–497.

Jones RM, Somerville LH, Li J, Ruberry EJ, Libby V, Glover G, Voss HU, Ballon DJ, Casey BJ (2011) Behavioral and neural properties of social reinforcement learning. J Neurosci 31:13039–13045.

Jones SH, Hemsley D, Ball S, Serra A (1997) Disruption of the Kamin blocking effect in schizophrenia and in normal subjects following amphetamine. Behav Brain Res 88:103–114.

Kamin L (1969) Predictability, surprise, attention, and conditioning. In: Punishment and aversive behavior (Campbell BA, Church RM, eds), pp 279–296. New York: Appleton-Century-Crofts.

Kant I (1965) Moral Law: Groundwork of the Metaphysics of Morals. London ; New York: Routledge.

Knutson B, Adams CM, Fong GW, Hommer D (2001a) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J Neurosci 21:RC159.

Knutson B, Fong GW, Adams CM, Varner JL, Hommer D (2001b) Dissociation of reward anticipation and outcome with event-related fMRI. Neuroreport 12:3683–3687.

Koritzky G, He Q, Xue G, Wong S, Xiao L, Bechara A (2013) Processing of time within the prefrontal cortex: recent time engages posterior areas whereas distant time engages anterior areas. Neuroimage 72:280–286.

Kosslyn SM, Cacioppo JT, Davidson RJ, Hugdahl K, Lovallo WR, Spiegel D, Rose R (2002) Bridging psychology and biology. The analysis of individuals in groups. Am Psychol 57:341–351.

Krienen FM, Tu P-C, Buckner RL (2010) Clan mentality: evidence that the medial prefrontal cortex responds to close others. J Neurosci 30:13906–13915.

Kuss K, Falk A, Trautner P, Elger CE, Weber B, Fliessbach K (2013) A reward prediction error for charitable donations reveals outcome orientation of donators. Soc Cogn Affect Neurosci 8:216–223.

Langleben DD, Loughead JW, Bilker WB, Ruparel K, Childress AR, Busch SI, Gur RC (2005) Telling truth from lie in individual subjects with fast event-related fMRI. Hum Brain Mapp 26:262–272.

Le Pelley ME, Oakeshott SM, McLaren IPL (2005) Blocking and unblocking in human causal learning. J Exp Psychol Anim Behav Process 31:56–70.

Lin A, Adolphs R, Rangel A (2012) Social and monetary reward learning engage overlapping neural substrates. Soc Cogn Affect Neurosci 7:274–281.

McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346.

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G (2011) Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci 31:2700–2705.

Mead NL, Baumeister RF, Gino F, Schweitzer ME, Ariely D (2009) Too tired to tell the truth: self-control resource depletion and dishonesty. J Exp Soc Psychol 45:594–597.

Mikhail J (2007) Universal moral grammar: theory, evidence and the future. Trends Cogn Sci (Regul Ed) 11:143–152.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Mill JS (1861) Utilitarianism (Crisp R, ed). Oxford ; New York: Oxford University Press.

Mitchell JP (2009) Social psychology as a natural kind. Trends Cogn Sci 13:246–251.

Mobbs D, Yu R, Meyer M, Passamonti L, Seymour B, Calder AJ, Schweizer S, Frith CD, Dalgleish T (2009) A key role for similarity in vicarious reward. Science 324:900.

Morgan TJH, Rendell LE, Ehn M, Hoppitt W, Laland KN (2012) The evolutionary basis of human social learning. Philos Trans R Soc Lond B Biol Sci 279:653–662.

Oades RD, Zimmermann B, Eggers C (1996) Conditioned blocking in patients with paranoid, non-paranoid psychosis or obsessive compulsive disorder: associations with symptoms, personality and monoamine metabolism. J Psychiatr Res 30:369–390.

Ochsner KN, Knierim K, Ludlow DH, Hanelin J, Ramachandran T, Glover G, Mackey SC (2004) Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. J Cogn Neurosci 16:1746–1772.

O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. Nat Neurosci 4:95–102.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. Neuron 33:815–826.

Pandya DN, Yeterian EH (1990) Prefrontal cortex in relation to other cortical areas in rhesus monkey: architecture and connections. Prog Brain Res 85:63–94.

Pavlov IP (1927) Conditioned reflexes. London: Oxford UP.

Penn DL, Sanna LJ, Roberts DL (2008) Social cognition in schizophrenia: an overview. Schizophr Bull 34:408–411.

Prados J (2011) Blocking and overshadowing in human geometry learning. J Exp Psychol Anim Behav Process 37:121–126.

Rak N, Bellebaum C, Thoma P (2013) Empathy and feedback processing in active and observational learning. Cogn Affect Behav Neurosci 13:869–884.

Rendell L, Fogarty L, Hoppitt WJE, Morgan TJH, Webster MM, Laland KN (2011) Cognitive culture: theoretical and empirical insights into social learning strategies. Trends Cogn Sci 15:68–76.

Rescorla RA (1999) Learning about qualitatively different outcomes during a blocking procedure. Anim Learn Behav 27:140–151.

Rescorla R, Wagner A (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II: Current Research and Theory (Black A, Prokasy W, eds), pp 64–99. New York: Appleton-Century-Crofts.

Romo R, Schultz W (1990) Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J Neurophysiol 63:592–606.

Ruff CC, Fehr E (2014) The neurobiology of rewards and values in social decision making. Nat Rev Neurosci 15:549–562.

Saxe R (2006) Uniquely human social cognition. Curr Opin Neurobiol 16:235–239.

Schoenbaum G, Chiba AA, Gallagher M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. Nat Neurosci 1:155–159.

Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80:1–27.

Schultz W, Apicella P, Ljungberg T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci 13:900–913.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.

Shalvi S, Eldar O, Bereby-Meyer Y (2012) Honesty requires time (and lack of justifications). Psychol Sci 23:1264–1270.

Shenhav A, Greene JD (2010) Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron 67:667–677.

Sip KE, Lynge M, Wallentin M, McGregor WB, Frith CD, Roepstorff A (2010) The production and detection of deception in an interactive game. Neuropsychologia 48:3619–3626.

Sip KE, Roepstorff A, McGregor W, Frith CD (2008) Detecting deception: the scope and limits. Trends Cogn Sci 12:48–53.

Souza MJ, Donohue SE, Bunge SA (2009) Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. Neuroimage 46:299–307.

Spence SA, Farrow TF, Herford AE, Wilkinson ID, Zheng Y, Woodruff PW (2001) Behavioural and functional anatomical correlates of deception in humans. Neuroreport 12:2849–2853.

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH (2013) A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16:966–973.

Sutton R, Barto A (1990) Time-Derivative Models of Pavlovian Reinforcement. In: Learning and Computational Neuroscience: Foundations of Adaptive Networks (Gabriel M, Moore J, eds), pp 497–537. MIT Press.

Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Haruno M, Cheng K, Nakahara H (2012) Learning to simulate others' decisions. Neuron 74:1125–1137.

Tanner C, Ryf B, Hanselmann M (2009) Geschützte Werte Skala (GWS): Konstruktion und Validierung eines Messinstrumentes (Protected values measure: construction and first validation of an instrument to assess protected values). Diagnostica 55:174–183.

Thompson-Schill SL, Braver TS, Jonides J (2005) Individual differences. Cogn Affect Behav Neurosci 5:115–116.

Thorndike EL (1911) Animal intelligence: Experimental studies. New York: MacMillan.

Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2006) Human neural learning depends on reward prediction errors in the blocking paradigm. J Neurophysiol 95:301–310.

Tremblay L, Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. Nature 398:704–708.

Van Horn JD, Grafton ST, Miller MB (2008) Individual variability in brain activity: a nuisance or an opportunity? Brain Imaging Behav 2:327–334.

Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412:43–48.

Watkins CJCH, Dayan P (1992) Technical note: Q-learning. Mach Learn 8:279–292.

Young L, Dungan J (2012) Where in the brain is morality? Everywhere and maybe nowhere. Soc Neurosci 7:1–10.

Appendix

A Appendix to Study 1

Social Cognitive and Affective Neuroscience Advance Access published November 3, 2014

doi:10.1093/scan/nsu130 SCAN (2014) 1 of 9

Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex

Azade Seid-Fatemi and Philippe N. Tobler
Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Zurich, Switzerland

When we are learning to associate novel cues with outcomes, learning is more efficient if we take advantage of previously learned associations and thereby avoid redundant learning. The blocking effect represents this sort of efficiency mechanism and refers to the phenomenon in which a novel stimulus is blocked from learning when it is associated with a fully predicted outcome. Although there is sufficient evidence that this effect manifests itself when individuals learn about their own rewards, it remains unclear whether it also does when they learn about others' rewards. We employed behavioral and neuroimaging methods to address this question. We demonstrate that blocking does indeed occur in the social domain and it does so to a similar degree as observed in the individual domain. On the neural level, activations in the medial prefrontal cortex (mPFC) show a specific contribution to blocking and learning-related prediction errors in the social domain. These findings suggest that the efficiency principle that applies to reward learning in the individual domain also applies to that in the social domain, with the mPFC playing a central role in implementing it.

Keywords: learning theory; social neuroscience; reward; cue competition; medial prefrontal cortex

INTRODUCTION

Learning to use environmental cues to predict upcoming events is crucial for adaptive behavior and reward-directed learning. It enables rapid response preparation and execution, thereby potentially providing an advantage over competitors. However, it also requires the use of limited cognitive resources that would be conserved if we prevented its occurrence whenever little or no new information was available. Thus, it would be more efficient if learning only took place when necessary, that is, when previous learning did not make it superfluous. For example, it is efficient not to learn a second cause when a first cause already fully explains an outcome. Associative learning research terms this the blocking effect because a previously learned stimulus that predictably leads to the same outcome "blocks" learning about a concurrently appearing novel stimulus. The underlying rationale is that learning to predict an outcome on the basis of a novel stimulus is redundant if we can already do so on the basis of a previously learned stimulus. In contrast, if the novel stimulus appears concurrently with a stimulus that does not predict the reward, then learning about the novel stimulus is not redundant and blocking does not occur. The blocking effect is expressed in a diminished behavioral response to novel stimuli that provide only redundant information about reward occurrence as compared with novel stimuli that provide non-redundant information. In this sense, the blocking effect represents efficient learning by definition. It should be noted, however, that this does not imply that blocking is also adaptive by definition (indeed coding of redundant information could be adaptive, e.g. when the system is irreducibly noisy). Efficient blocking of novel learning has been observed in different species including rats (Kamin 1969), monkeys (Waelti *et al.*, 2001) and humans (Tobler *et al.*, 2006; Prados 2011; Eippert *et al.*, 2012).

Received 25 February 2014; Revised 1 September 2014; Accepted 14 October 2014

The authors thank several members of the Department of Economics, University of Zurich, and Bertram Gerber for fruitful discussions, JiSoo Park for assistance with data collection, and Christopher Burke, Tamara Herz, Quentin Huys and Yosuke Morishima for helpful comments on a previous version of this article. This work was supported by funding from the Swiss National Science Foundation (PP00P1_128574). The authors also acknowledge the Neuroscience Center Zurich (ZNZ) and the Zurich Center for Integrative Human Physiology (ZIHP).

Correspondence should be addressed to Azade Seid-Fatemi, Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland.
E-mail: azade.seid-fatemi@econ.uzh.ch.

The blocking effect has been investigated primarily in individuals learning about rewards for themselves and has been captured by formal models of learning in the individual domain (Rescorla and Wagner 1972). However, sometimes rewards are not received by us, but by others. Cues that are redundant for others might nevertheless be of value to us as they may provide us with relevant information that is either relevant now or will become relevant at a later point in time. In particular, these cues may enter different associations with rewards for ourselves as compared with rewards for others. Therefore, in a social context, we might want to keep track of cues concerning rewards for others, even if they are redundant for others. By extension, whether we still rely on previous learning (resulting in blocking) or track all information available to us in the environment (not resulting in blocking) remains an open question for reward learning in the social domain. Indeed, at least with some forms of social learning, such as socially transmitted food preferences, there appears to be little blocking (Galef and Durlach 1993). In this study, we investigate whether and, if so, how the efficiency principle represented by the blocking effect extends to reward learning in the social domain.

On the neural level, the blocking effect has only been investigated during reward learning in the individual domain. It is expressed in a reduced neural response to blocked stimuli as compared with non-blocked (control) stimuli. This effect has been observed in the ventromedial prefrontal cortex (vmPFC) and the striatum using juice rewards (Tobler *et al.*, 2006) as well as in the amygdala using electric shocks (Eippert *et al.*, 2012). Due to their potential usefulness to oneself, learning about rewards received by others may to a certain extent engage the same neural regions that process rewards received by oneself. Previous work suggests that, in particular, ventral and medial regions of PFC are engaged during reward learning in a social context (Behrens *et al.*, 2008; Burke *et al.*, 2010; Suzuki *et al.*, 2012; Zhu *et al.*, 2012). Thus, if the blocking effect indeed occurred in the social domain as well, the question arises whether it would be implemented by the same vmPFC regions that implement it in the individual domain. Alternatively, specific mPFC regions may be involved in blocking redundant reward learning in the social domain. To investigate these questions, we developed a social variant of the blocking paradigm and used functional magnetic resonance imaging (fMRI) to examine the role of the mPFC in learning to predict monetary rewards.

MATERIALS AND METHODS

Participants

Thirty-eight participants (17 female; aged 21.0 ± 0.4 years; range: 18–28) took part in this study. None of the participants had prior histories of neurological or psychiatric disorders and all had normal or corrected-to-normal vision. Written informed consent was obtained from all participants, and the study was approved by the Research Ethics Committee of the Canton of Zurich.

Experimental design

We investigated the blocking effect (Kamin 1969) with respect to monetary rewards in the social and individual domains in two separate conditions. Individual rewards were received by the participant, social rewards by another person. Two female volunteers served as the other person in the social condition. We used two existing persons as volunteers who received money at the end of the experiment to ensure that the consequences of the social rewards were as real as those of the individual rewards. Moreover, using two rather than only one volunteer served to make the social conditions of the experiment more engaging (i.e. more varied and less monotonous) and thereby prevent adaptation. The participants never met the two volunteers face-to-face, but read a brief description of them before the experiment began, which included their initials and information about their gender and age. Before the experiment, we determined the amount of the individual reward according to the social preferences of each subject. We did this to ensure that the rewards in the social and the individual condition had the same subjective value. To achieve this, we used a variant of the Becker–DeGroot–Marschak method (BDM; Becker *et al.*, 1964). Specifically, before the experiment, we asked subjects to indicate the amount of money (in Swiss Francs, CHF; between CHF 1 and 100) that, if delivered to them, was as valuable as delivering CHF 60 to the other person. The amount of CHF 60 was chosen based on pilot studies with a separate set of subjects showing that CHF 60 yielded affordable individual equivalence amounts (CHF 45.80 ± 1.90). The bid was then compared with a random number between 1 and 100 generated by the computer. If the number was greater than or equal to the subject's bid, they received the indicated amount of money. If the number was lower than the bid, they received nothing and the other person received CHF 60. Thus, the procedure provided an incentive-compatible way of obtaining individual reward amounts that corresponded to the value of social reward amounts. The outcome of the procedure had no influence on the payout or the number of rewards, the participants gained from the actual experiment. The bid was obtained before the experiment and used to set the individual reward amount in the experiment such that it had the same value as the social reward amount, given the subject's social preferences.

As previously described (Waelti *et al.*, 2001; Tobler *et al.*, 2006), the within-subject version of the blocking procedure comprised three consecutive phases. In each of these phases, participants were presented with visual stimuli that were associated with the delivery of different social or individual outcomes (Figure 1A). All of the stimuli used were abstract colored shapes presented on a white background and were similar to those used in previous blocking experiments (Waelti *et al.*, 2001; Tobler *et al.*, 2006). Each trial had either a social or individual outcome, never both, and different stimuli were used for each of the conditions (and thus recipients).

In the first (pretraining) phase, the A (experimental) stimuli (A_{SOCIAL} and $A_{INDIVIDUAL}$) were paired with a social or individual reward. In contrast, the B (control) stimuli (B_{SOCIAL} and $B_{INDIVIDUAL}$) were not paired with a reward. Stimuli were presented 20 times (see Supplementary Table S1) each and the identities of the stimuli were counterbalanced across participants. Each trial started with a 4-s intertrial interval (ITI) that varied from 2 to 6 s (Figure 1B). Stimuli were presented for 1.5 s at random either to the left or the right of the fixation cross. The outcome was presented concurrently with the stimulus for another 1.5 s. During the presentation of any given stimulus, the participants were to perform a specific key press corresponding to the recipient and to the outcome that would follow the stimulus. In particular, upon each stimulus presentation, participants had to indicate whether they expected reward for self, no reward for self, reward for others or no reward for others by pressing a key with the index or middle finger of their left or right hand. Thus, there was an individual and social reward key and an individual and social no-reward key, and participants were asked to press one of these keys in each trial. This allowed us to measure recipient- as well as outcome-specific learning. Condition-to-hand (individual or social) and key-to-reward (reward or no reward) assignments were counterbalanced across participants. Trials in which the participant failed to respond or responded too late were repeated later. Visual stimuli as well as response recordings were controlled using Cogent 2000 (Wellcome Department of Imaging Neuroscience, London, UK) as implemented in Matlab.

fMRI scanning started in the second (compound conditioning) phase. Visual stimuli were presented on a display that participants viewed via a mirror fitted to the top of the head coil. In the compound phase, A stimuli were presented together with X stimuli (X_{SOCIAL} and $X_{INDIVIDUAL}$), forming rewarded compounds. As a control, B stimuli were presented together with Y stimuli (Y_{SOCIAL} and $Y_{INDIVIDUAL}$) and also followed by a reward. In AX trials, the upcoming reward was predicted by the A stimuli. Therefore, efficiency considerations would suggest that the X stimuli should be blocked from learning. Specifically, in formal learning theory (Rescorla and Wagner 1972), the blocking effect is explained by "cue competition" and operationalized by summing up the associative strengths (predictive values) of all stimuli present in a given trial. In early AX trials, the sum was already close to the value of the reward itself, leading to little difference between predicted and actual value when the reward occurs. In contrast, in early BY trials, the B stimuli did not predict a reward and the summed associative strength was low. Thus, cue competition explains why Y stimuli, but not X stimuli, should be learned as reward-predicting stimuli. AX and BY trials were presented in 24 trials per condition and intermixed with 14 A and B trials per condition, which served to maintain the previously learned associations (see Supplementary Table S1). To prevent compound trials in general from being associated with reward, we included control compound trials (CZ trials) that were unrewarded (12 trials each for the social and individual condition; see Supplementary Table S1). These unrewarded, occasionally interleaved compound stimuli are not standardly used in the blocking procedure. We used them in order to ensure that participants paid attention to each of the individual stimuli that constituted a compound rather than automatically associating the co-occurrence of any two stimuli with reward. We thereby aimed to keep learning more elemental than configural (Melchers *et al.*, 2008).

In a third phase, X and Y stimuli were presented alone in unrewarded test trials. Under the assumption that previous learning blocks subsequent learning, the X stimuli should have been blocked from being associated with social or individual reward, whereas the Y stimuli should have been associated with reward. Y and X trials were presented in 14 trials each and randomly intermixed with A and B trials (14 trials), AX and BY trials (24 trials) and control compound trials (12 trials), again, to maintain previously learned associations (see Supplementary Table S1). As before, A, AX and BY trials were followed by reward in order to maintain the previously learned associations.

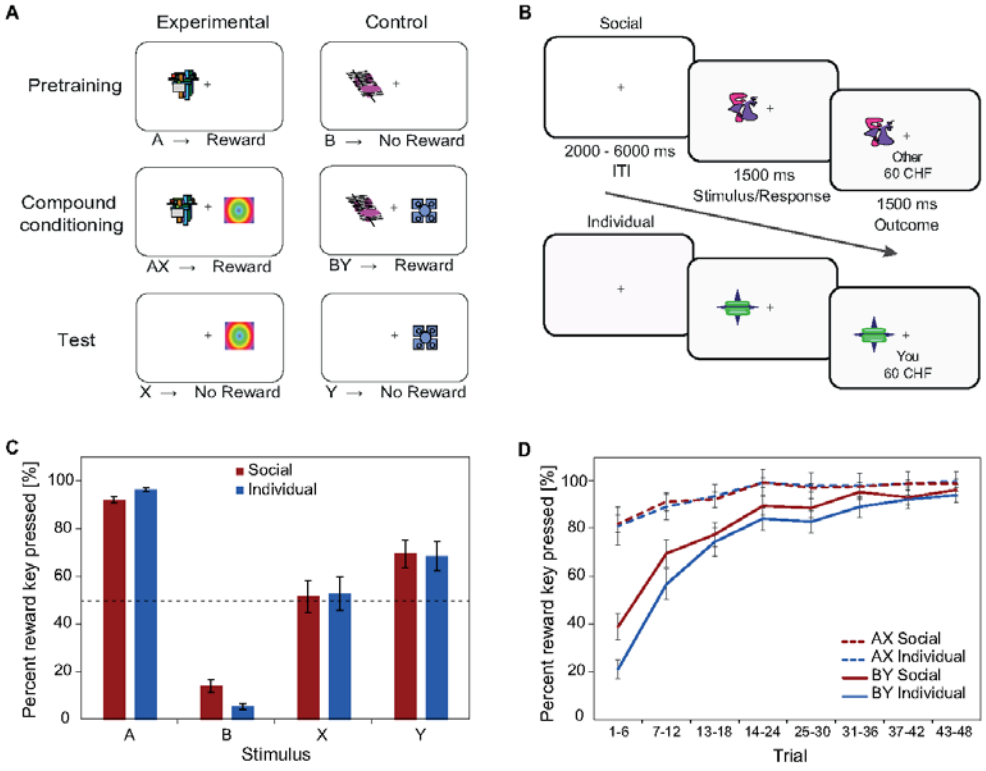

Fig. 1 Experimental design and behavioral results. (A) Three phases of blocking paradigm with monetary rewards. During pretraining, participants learned to associate stimuli with the presence or absence of monetary outcomes. Reward-predicting A stimuli were followed by a monetary reward, but not neutral B stimuli (each of these had a social and an individual variant, see Figure 1B). During compound conditioning, X and Y stimuli appeared together with A and B stimuli in rewarded compounds. In AX trials, the reward was fully predicted by the A stimulus. Therefore, the X stimulus was expected to be blocked from learning. In contrast, in BY trials, the reward was not predicted by the B stimulus, so the Y (control) stimulus was expected to be learned as a reward-predicting stimulus. During the test phase, X and Y stimuli were presented alone and remained unrewarded. Although learned stimulus Y was expected to predict upcoming reward, blocked stimulus X was not. During the compound and the test phase, trial types of the previous phases were also presented in order to maintain learned associations. (B) Example of pretraining trials. Abstract visual stimuli were presented in random order, either to the left or the right of the fixation cross. Upon presentation of a stimulus, the participants were to perform a specific key press corresponding to the recipient (self or other) and to the outcome (reward or no reward) that would follow the stimulus. The outcome was shown together with the stimulus for another 1.5 s. The ITI varied between 2 and 6 s. (C) Participants showed an increase in reward-expecting responses (quantified as percentage of key presses) to reward-predicting A stimuli and Y (control) stimuli as compared with unrewarded B stimuli and blocked X stimuli in the social as well as the individual condition, suggesting blocking effects in both cases. Key presses for A and B are shown for trials from all three phases. Error bars indicate SEMs. (D) Mean learning curves averaged across participants showed a stronger increase in the BY as compared with the AX condition for the social as well as the individual condition. Shown is the percentage of reward key presses over time (each bin averaged over six trials). Error bars indicate SEMs.

During fMRI scanning, the experiment was split into three sessions that did not coincide with conditioning phases in order to prevent rapid extinction of Y stimuli during test trials. The compound conditioning phase spanned the first scanning session and the first half of the second scanning session. The test phase began with the second half of the second scanning session and ended at the end of the third scanning session.

Participants were instructed that, at the end of the experiment, a portion of the rewards accumulated in correctly predicted trials would be paid out to them and the other two individuals, respectively. To ensure that everyone received approximately the same amount irrespective of their bid in the BDM, we adjusted the percentage for each participant individually. To keep them engaged throughout the task, in each trial in which participants failed to respond or responded too slowly, CHF 1 was deducted from their final monetary payment and the three participants with the highest number of correct responses received an additional payment (CHF 20). The highest percentage of trials missed by a participant was 3%.

Social and individual reward expectation was defined as the percentage of the social and individual reward key pressed, respectively. Reward expectations were evaluated using paired *t*-tests and two-way repeated-measures analysis of variances. The degree of participant-specific behavioral blocking was calculated as the difference between recipient-specific reward key presses for Y stimuli and those for X stimuli. The larger the difference, the stronger the blocking effect. Comparing the responses to Y with those to X stimuli is the standard approach to determining whether blocking has taken place. However, for the neural data, we also analyzed outcome-related activation in AX and BY trials to measure the differential learning responses when comparatively small or large amounts of learning occur, respectively (see above and below).

fMRI data acquisition

fMRI data were acquired on a Philips Achieva 3 T whole-body scanner equipped with an eight-channel head coil (Philips Medical Systems, Best, The Netherlands) at the Laboratory for Social and Neural Systems

Research, University of Zurich. We acquired gradient-echo T2*-weighted echoplanar images (EPIs) with blood-oxygen-level-dependent contrast (slices/volume, 33; repetition time, 1.75 s). Approximately 530–710 volumes were collected per session (variation was due to experiment phase and individual differences in the number of repeated trials) along with five "dummy" volumes at the start of the scanning session to allow for magnetization to stabilize to a steady state. Scan onset times varied relative to stimulus onset times. Slice orientation was tilted 20° away from the anterior commissure-posterior commissure line, caudal > rostral. Imaging parameters were: echo time, 30 ms; field-of-view, 240 mm; in-plane resolution, 3 mm; slice thickness, 3 mm; interslice gap, 0.75 mm. A T1-weighted structural image was also acquired for each participant. These high-resolution T1-weighted structural scans were coregistered to their mean EPIs and averaged to permit anatomical localization of the functional activations at the group level.

fMRI data analysis

fMRI data processing and statistical analyses were carried out using statistical parametric mapping (SPM8; Wellcome Department of Imaging Neuroscience, London, UK). Data preprocessing consisted of realignment, coregistration, segmentation, spatial normalization using the DARTEL toolbox and smoothing using a Gaussian kernel with a full width at half maximum of 10 mm. Data analysis was performed using a general linear model approach. The first-level design matrix of each participant included separate regressors for each of the four learned stimulus conditions ($A_{INDIVIDUAL}$, A_{SOCIAL}, $B_{INDIVIDUAL}$, B_{SOCIAL}) modeled at the event-onset time, the compound conditioning trials at the time of the outcome ($AX_{INDIVIDUAL}$, AX_{SOCIAL}, $BY_{INDIVIDUAL}$, BY_{SOCIAL}) to capture prediction error-related responses during learning, and the four test trial types modelled at the event-onset time ($X_{INDIVIDUAL}$, X_{SOCIAL}, $Y_{INDIVIDUAL}$, Y_{SOCIAL}) to capture blocking. In order to identify brain regions that correlate with prediction error during compound learning trials, we parametrically modulated the AX and BY regressors with trial-wise and mean-corrected prediction errors (δ) derived from a standard reinforcement learning model (see below). To account for the variance that can be explained by stimulus presentation, we created two additional regressors for compound conditioning trials at event-onset time, combining AX and BY trials into a single regressor (AX/BY_{SOCIAL}, $AX/BY_{INDIVIDUAL}$). Finally, we included regressors of no interest for the unrewarded compound trials and for participant-specific movement parameters (three regressors for rotation and three for translation). All regressors were convolved with the canonical hemodynamic response function. For each regressor, we included all trials irrespective of the participants' response.

In order to identify brain regions involved in prediction error-based learning during compound conditioning, we parametrically modulated the AX and BY regressors with mean-corrected prediction errors derived from a variant of a simple reinforcement learning model (Rescorla and Wagner 1972). In each trial, prediction errors were computed according to $\delta_t = \alpha\ (\lambda_t - V_t)$, where V_t corresponds to the value V predicted by all stimuli presented in trial t, λ_t corresponds to the reward in trial t, and α corresponds to the learning rate. The learning rate determines how much weight is given to recent experience as captured by the prediction error. It is a free parameter that can be used to characterize how quickly participants learn in different conditions (see e.g. Burke *et al.*, 2010). We estimated the learning rate by fitting the prediction error model above to the trial-by-trial percentage of reward keypress responses in BY trials, averaged across participants. These keypresses are a measure of participants' reward prediction in a given trial. As different keys were used for the prediction of social and individual reward, we were able to estimate separate learning rates by using their keypresses for the social and individual condition, respectively. The estimated learning rate was 0.10 for the social condition and 0.15 for the individual condition (no significant difference).

The prediction error in a given trial is used to update the associative strengths of all stimuli present in that trial. For example, in the initial BY trials, the associative strength of BY is low, so a reward should generate a positive prediction error. During training with the BY compound, the reward becomes more and more predictable and, according to theory, the prediction error gradually decreases (Supplementary Figure S1). On the other hand, in the initial AX trials, the associative strength of AX is already high (i.e. reward is already fully predicted) due to the pretrained A stimuli, so a decrease in prediction error should not occur. Taken together, for regions involved in learning, over the course of compound conditioning, we expect to observe a greater reduction in prediction-error-related activity in BY than in AX trials. This would be captured by a better fit with a parametric modulator that models a decreasing prediction error signal in BY compared with AX trials.

Linear contrasts of regression coefficients of A *vs* B (stimulus response), BY *vs* AX (prediction error modulator), and Y *vs* X (stimulus response) were computed at the single-participant level and then taken to group-level analyses where we used one-sample *t*-tests or correlations with participant-specific degree of blocking in the social or individual domain. Correction for multiple comparisons (familiywise error, FWE; $P < 0.05$) was performed either in areas of interest or at the whole-brain level. Our a priori region of interest for the individual condition was defined functionally as a 15-mm sphere around the peak of a previously reported coordinate reflecting individual blocking in vmPFC (−18, 36, −6) (Tobler *et al.*, 2006). Moreover, we assessed blocking effects within spheres around peak activations identified by the independent reward expectation contrast of A *vs* B. Outside these regions of interest, correction for multiple comparisons was performed at the whole-brain cluster level ($P < 0.05$, cluster-inducing threshold: $P < 0.001$). In the figures, the left side of the brain is shown on the left.

RESULTS

Behavioral results

We used a within-subject design that employed three phases to test blocking in the social and individual domain (Figure 1A and B, see Materials and Methods). In the pre-training phase, participants learned to associate one stimulus (A) with reward and another (B) with no reward; they maintained these associations in the subsequent compound and test phases. The association between A and reward was expected to block individual learning in the compound phase, but it was an open question whether this would also hold for social learning. The degree to which blocking had occurred was assessed in unrewarded trials in the test phase. First, we tested whether participants had learned the (previous) stimulus-outcome associations: participants predicted recipient-specific reward outcomes when presented with reward-predicting A stimuli, but not when presented with B (control) stimuli (Figure 1C). This resulted in a significantly higher number of reward key presses for A *vs* B stimuli (over all phases), for both social and individual conditions (social: $92.0 \pm 1.3\%$ *vs* $14.1 \pm 2.4\%$, $t_{(37)} = 23.75$; individual: $96.5 \pm 0.7\%$ *vs* $5.3 \pm 1.4\%$ (mean ± SEM), $t_{(37)} = 47.34$, both $P < 0.001$). The results were very similar when the analysis was limited to A and B trials of the pretraining phase (social: $88.2 \pm 2.0\%$ *vs* $17.0 \pm 3.3\%$, $t_{(37)} = 15.08$; individual: $95.1 \pm 1.3\%$ *vs* $6.9 \pm 2.4\%$ (mean ± SEM), $t_{(37)} = 26.23$, both $P < 0.001$), indicating that the pretraining phase, which took place before the compound conditioning phase, was successful in both the social and the individual

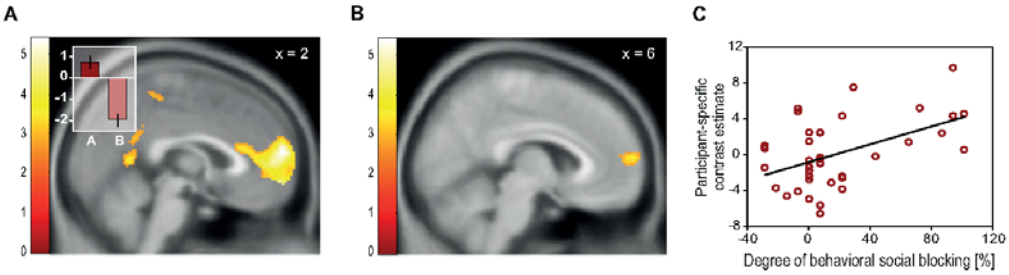

Fig. 2 Activity in mPFC reflects expectation and blocking of social rewards. (A) mPFC responses were higher to reward-predicting A stimuli as compared with neutral B stimuli (2, 60, 16; $P < 0.05$, FWE-corrected). Contrast estimates (inset) show mPFC responses to A and B stimuli separately. Error bars indicate SEM. (B and C) Differences in activation responses to Y as compared with X increased in the mPFC (6, 60, 12, $P < 0.05$, FWE small-volume corrected) with degree of blocking in the social condition. Blocking was quantified as the difference in reward-expecting responses to non-blocked stimulus Y compared with blocked stimulus X. Color bars indicate *z*-scores.

condition. Taken together, the participants learned to discriminate in an outcome- and recipient-specific manner between stimuli predicting reward and stimuli predicting no reward.

Next, we investigated whether the course of learning differed in the BY *vs* AX condition. In the BY condition, reward key presses gradually increased in both the social and the individual condition (Figure 1D). Thus, in early trials, participants showed only a low number of reward key presses, but learned the association over time. In contrast, there was only a very mild increase in reward key presses in the AX condition as participants were already predicting the reward outcome for AX at the beginning of the compound phase. To assess whether the increase in reward key presses in BY trials differed from that in AX trials, we compared the first six trials (early) with the last six trials (late) and found that, in both the social and the individual condition, trial type (BY *vs* AX) interacted with time (early *vs* late compound trials; social: $F_{(1,37)} = 54.82$; individual: $F_{(1,37)} = 257.34$; both $P < 0.001$), indicating that more learning occurred during BY trials than during AX trials.

We then assessed whether the blocking effect manifests itself not only in the individual but also in the social domain by comparing participants' reward expectation to potentially blocked X stimuli *vs* Y (control) stimuli in non-rewarded test trials. There was an increase in reward key presses for Y (control) stimuli as compared with X stimuli in both the social ($t_{(37)} = 3.07$, $P < 0.005$) and the individual ($t_{(37)} = 2.48$, $P < 0.05$) condition (Figure 1C). Although the difference between Y and X was smaller than that between A and B (social: $t_{(37)} = 8.76$; individual: $t_{(37)} = 11.06$, both $P < 0.001$), participants clearly treated Y and X differently in the individual as well as the social condition. If participants had failed to learn anything about the outcome of a given stimulus, we would expect performance at chance level, as there was a 50% chance of pressing either the reward or the no-reward key (dotted lines in Figure 1C). For the social as well as the individual condition, reward key presses for X stimuli did not differ from chance (social: $51.5 \pm 6.8\%$, $t_{(37)} = 0.22$, $P = 0.83$; individual: $52.8 \pm 6.8\%$, $t_{(37)} = 0.41$, $P = 0.68$) whereas those for Y (control) stimuli occurred more often than 50% (social: $69.7 \pm 5.8\%$, $t_{(37)} = 3.40$; individual: $68.4 \pm 6.2\%$, $t_{(37)} = 2.97$, both $P < 0.05$), again confirming that blocking had occurred in the individual as well as in the social condition.

To measure participant-specific differences in the blocking effect and compare social and individual blocking, we determined each participant's degree of blocking by calculating the difference between reward key presses for Y and X. Interestingly, the degree of blocking was similar for the social and individual condition ($t_{(37)} = 0.53$, $P = 0.6$). Moreover, across participants, the degree of blocking in the social condition was correlated with the degree of blocking in the individual condition ($R^2 = 0.45$, $P < 0.001$). Thus, from a behavioral standpoint, blocking in the social and individual conditions were related.

We also tested whether the social and individual conditions differed in their salience as assessed by differences between the conditions with respect to response time. There were no significant response time differences between social and individual test trials (X: 853.8 ± 14.9 ms *vs* 852.6 ± 15.6, $t_{(37)} = 0.10$, $P = 0.92$; Y: 823.0 ± 16.2 ms *vs* 839.4 ± 18.6, $t_{(37)} = -0.81$, $P = 0.42$). In A and B trials, participants responded faster in the individual as compared with the social condition (A: 729.2 ± 9.0 ms *vs* 785.3 ± 11.5 ms, $t_{(37)} = -6.04$; B: 748.0 ± 10.6 ms *vs* 790.8 ± 10.5 ms, $t_{(37)} = -5.50$; both $P < 0.001$). Thus, while individual A and B trials may have been more salient than social A and B trials, there is no evidence for a difference in salience between social and individual Y and X trials, with which blocking was assessed at the neural level.

fMRI results

Blocking in the social domain

First, we investigated responses in A *vs* B trials with respect to the participants' expectation that another person would receive a monetary reward (social condition). We found stronger activation in the mPFC when participants expected another person to receive a reward than when they did not (Figure 2A; 2, 60, 16; $t_{(37)} = 5.41$; $P < 0.05$, whole-brain FWE cluster-level corrected; see Table 1 for additional whole-brain corrected activations).

To test whether the mPFC activity found for A *vs* B shows a blocking effect as well, we investigated the contrast of Y *vs* X. This is the standard contrast (Tobler *et al.*, 2006; Eippert *et al.*, 2012) used to test for the blocking effect as it captures reduced neural responses to the (blocked) X stimulus as compared with the (non-blocked) Y stimulus. An increased response to the Y compared with the X stimulus reflects the stronger reward prediction for Y *vs* X, similar to the difference in reward prediction for A *vs* B, but uncontaminated by actual reward delivery. We therefore used a mask including a 10-mm sphere around the peak coordinate from the contrast of A *vs* B in the social condition and performed an independent second-level correlation analysis of differential brain activation in Y *vs* X against the participant-specific behavioral difference of Y *vs* X. We found a correlation in mPFC (Figure 2B and C; 6, 60, 12; $t_{(36)} = 3.63$; $P < 0.05$, FWE small-volume corrected). Thus, within mPFC, similar subregions showed activations reflecting reward expectation and blocking in the social domain.

Blocking in the individual domain

To confirm previous findings on blocking in the individual domain, we examined responses related to reward expectation and blocking in the individual condition. Specifically, we analyzed activity in the vmPFC, a region identified in a previous study on the blocking effect involving liquid reward (Tobler *et al.*, 2006). We found that activity in the vmPFC was stronger for reward-predicting stimuli than for neutral stimuli (Supplementary Figure S2A; −4, 40, −6; $t_{(37)} = 3.81$; $P < 0.05$, FWE small-volume corrected) and increased with the degree of behavioral blocking (Supplementary Figure S2B and C: −6, 42, −4; $t_{(36)} = 3.84$; $P < 0.05$, FWE small-volume corrected). These data suggest that vmPFC activations reflect both blocking and reward expectation in the individual condition.

Table 1 Brain regions exhibiting additional learning- or blocking-related activation

Brain region	*x*	*y*	*z*	*t*
A > B (social)				
mPFC	2	60	16	5.41
Posterior cingulate cortex	−6	−52	16	4.74
A > B (social *vs* individual)				
Rolandic operculum	−60	−6	10	5.32
Precuneus	8	−54	58	4.81
mPFC	8	56	12	4.52
Middle occipital gyrus	48	−78	20	4.35
Y > X (social *vs* individual)				
Lateral prefrontal cortex	26	60	6	4.36
BY_PM > AX_PM (social)				
Parietal cortex	34	−62	38	4.84
dmPFC	10	30	38	4.65

Regions that survive whole-brain FWE correction at the cluster level, with a cluster-inducing threshold of $P < 0.001$. Coordinates are denoted by *x*, *y*, *z* (in mm; MNI space).

Comparison of social and individual conditions

On the behavioral level, we found similar reward expectation and blocking effects for the social and individual conditions. Nevertheless, it is possible that distinct regions in the brain keep track of the reward recipient. We therefore tested whether the mPFC response reflecting blocking and reward expectation is stronger in the social than in the individual domain. First, we assessed whether activation for A *vs* B is specific for the social condition. The more dorsal part of the mPFC that was identified for social learning and blocking (Figure 2A and B) responded more strongly in the social than in the individual condition (A *vs* B social > A *vs* B individual: 8, 56, 12; $t_{(37)} = 4.52$; Figure 3A and B; $P < 0.05$, whole-brain FWE cluster corrected; additional whole-brain corrected activations are shown in Table 1). Moreover, activity in the same region was also stronger for blocking in the social than in the individual condition. In other words, we found activity in mPFC for the contrast of Y *vs* X social > Y *vs* X individual (12, 56, 10; $t_{(37)} = 3.57$; Figure 3C and D; $P < 0.05$, FWE small-volume corrected in a 10-mm sphere around the peak coordinate of A *vs* B social > A *vs* B individual). This activation extended into more lateral parts of the mPFC with an additional peak at 26, 60, 6 ($P < 0.05$, whole-brain FWE cluster-level corrected). Note that these preferential neural effects of blocking in the social condition occurred in the absence of significant behavioral or value differences between the individual and the social conditions.

Development of blocking in the social domain

After establishing that the mPFC plays a preferential role in blocking in the social domain, we investigated social learning during the compound conditioning phase. In the AX trials, the social reward was already fully predicted by the pretrained stimulus A and therefore the reward was expected to elicit little or no prediction error signal.

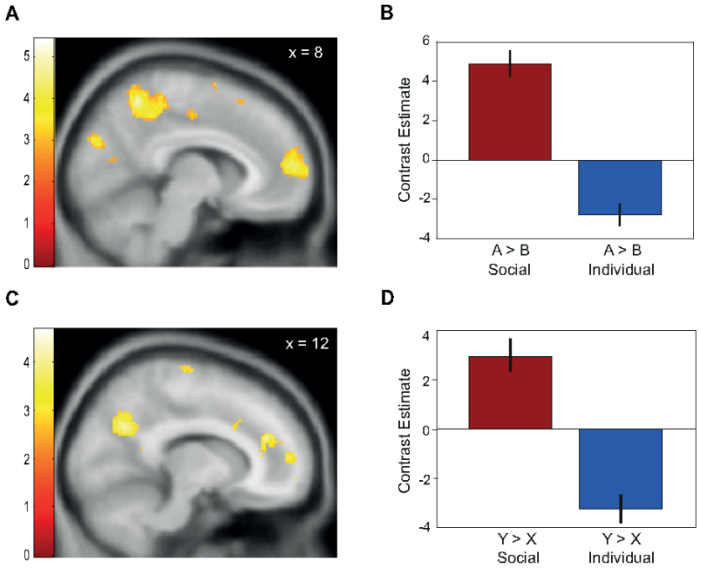

Fig. 3 Preferential responses to reward expectation and blocking in the social as compared with the individual condition. (A) and (C) Stronger responses in the mPFC in the social than in the individual condition for the contrast of A *vs* B (8, 56, 12, $P < 0.05$, FWE small-volume corrected) and Y *vs* X (12, 56, 10, $P < 0.05$, FWE small-volume corrected). (B and D) Contrast estimates show mPFC response for the contrast A *vs* B and Y *vs* X separately for the social and the individual condition. Error bars indicate SEM. Color bars indicate *z*-scores.

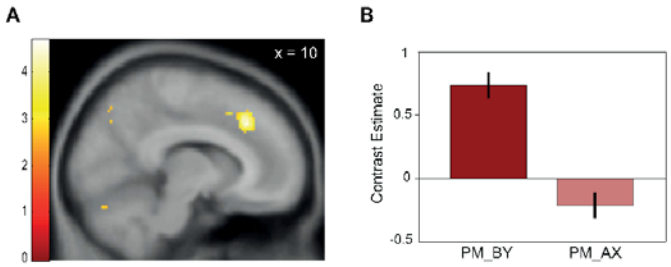

Fig. 4 Differential activations during the development of the blocking effect. (A) Activation in the dmPFC (peak at 10, 30, 38; $P < 0.05$, FWE cluster corrected) was better fitted by a parametric modulator that modeled decreases in prediction error in BY (PM_BY) as compared with AX trials (PM_AX). Color bar indicates *z*-score. (B) Contrast estimates show response in the dmPFC for the parametric modulators of BY and AX separately. Error bars indicate SEM.

In contrast, in the BY trials, social reward was not predicted as B had not been rewarded in pretraining. Consequently, the reward outcome was expected to generate a sizeable prediction error in early BY trials. As learning progressed from trial to trial, the reward outcome was expected to elicit a gradually decreasing prediction error. Accordingly, we tested for better fits with decreases in prediction error in BY compared with AX trials. We therefore parametrically modulated the AX and BY regressors in the social condition with trial-wise mean-corrected prediction errors derived from a simple reinforcement learning model (see Materials and Methods and Supplementary Figure S1). We found that activation in the dorsomedial PFC (dmPFC) fitted the parametric modulator for BY compared with AX trials better, reflecting the more strongly decreasing prediction error responses in BY trials in the social condition (Figure 4A and B; 10, 30, 38; $t_{(37)} = 4.65$; $P < 0.05$, FWE cluster corrected; see Table 1 for additional whole-brain corrected activation). Additionally, at less stringent statistical thresholds, we found that the differential dmPFC activation was more sensitive to prediction errors in the social as compared with the individual condition (BY *vs* AX social > BY *vs* AX individual: 12, 30, 40; $t_{(37)} = 3.22$; $P < 0.001$, uncorrected). Furthermore, the differential fit of the prediction-error-related activity in the dmPFC correlated with the degree of behavioral blocking in the social condition (−12, 26, 40; $t_{(36)} = 3.40$; $P < 0.001$, uncorrected). Although both findings should be interpreted with care due to their uncorrected nature, they are in line with the notion that the dmPFC preferentially codes prediction errors during reward learning in the social rather than the individual domain and that this prediction error coding is related to participant-specific differences in blocking in the social domain.

DISCUSSION

In this study, we investigated whether and, if so, how the efficiency principle represented by the blocking effect extends to reward learning in the social domain. Our behavioral results did indeed reveal a blocking effect in the social domain and thereby suggest that, as in the individual domain, efficiency is weighted more heavily than the complete encoding of all available information. Thus, the same mechanism that leads to efficient reward learning in the individual domain also serves as an efficient strategy to optimize learning in the social domain. Moreover, the degree to which the effect manifested itself in the two domains was correlated across participants. Nevertheless, although we found similar and correlated blocking effects in the two domains on the behavioral level, on the neural level, we found that the more dorsal mPFC assumes a preferential role for the blocking of socially relevant cues.

At the behavioral level, we found blocking not only with individual, but also with social learning. Thus, our study suggests that blocking occurs in at least some forms of social learning. This was not obvious from the outset as attempts to show blocking, for example, in the domain of socially transmitted food preferences were not successful (Galef and Durlach 1993). In the case of socially transmitted food preferences, however, the definition of the unconditioned stimulus (social interaction with a demonstrator rat) and its relationship to the dependent variable (food consumption by the observer rat) is less obvious than in more standard paradigms of individual learning. In contrast, we used similar response requirements and clearly defined rewards in both individual and social conditions, which facilitated the comparison of individual and social learning.

Blocking and reward expectation effects in the social domain were enhanced over and above those in the individual domain in relatively more dorsal regions of mPFC. This preferential relationship with social effects arose even though we equated subjective values and response requirements in the individual and social conditions (see Materials and Methods). Thus, we can exclude the possibility that the mPFC activation simply reflects differential values or response requirements related to one's own or others' rewards. It is not likely due to differences in salience either (e.g. Leathers and Olson 2012) as response times were similar in X and Y trials in the social and individual conditions. These conclusions are further supported by a control analysis: We obtained subjective desirability ratings indicating how much each participant cared about the two other people receiving a monetary payoff during the task. Including this variable as a covariate of no interest allowed us to identify a very similar increase in activation in mPFC that correlated with the degree of blocking observed in the social condition. Thus, the preferential contribution of the more dorsal mPFC to social learning and blocking cannot be explained by differential value- and salience-related effects.

Incidentally, the control analysis also renders an explanation in terms of conflict less likely: those who cared less about others receiving a reward should have felt more conflict when they had to perform a movement that would be followed by such a reward. However, the desirability ratings were not related to participant-specific differences in response times during social reward expectation or blocking. In line with the absence of a role of conflict, our activations occurred in more anterior and ventral locations of mPFC than those typically associated with conflict (Botvinick *et al.*, 2001; Kerns *et al.*, 2004; Shenhav *et al.*, 2013).

Our data suggest that the relatively dorsal mPFC regions contribute to other-directed reward learning by implementing an efficient

learning mechanism originally described in empirical studies and formal models of individual learning (Kamin 1969; Rescorla and Wagner 1972). Our data thereby converge with reports of relatively dorsal mPFC involvement in other aspects of social learning. For example, Behrens *et al.* (2008) found that activity in the dmPFC correlates with errors in the predicted helpfulness of confederate advice. In another case, dmPFC activation during an inspection game was found to correlate with the degree to which players thought they influenced their opponent's behavior (Hampton *et al.*, 2008). Thus, activity of the dorsal mPFC can be captured particularly well with formal models of social learning with the unifying explanation that this region encodes social reward prediction errors.

Responses in the dmPFC reflected the gradual decrease in prediction errors in BY trials, indicating that this region processes the change in prediction errors as anticipated by formal learning theory. Note that this dmPFC activation is more dorsal and posterior than the mPFC region we found to be sensitive to blocking in the social domain, suggesting that different subregions of the dmPFC are engaged at different stages of social learning. Future research may therefore focus on the mechanisms underlying the development of blocking in the social domain and investigate in more detail how the development of the effect in the compound phase relates to its expression in the test phase.

In our social condition, participants predicted whether another person would obtain a reward. Outcomes related to others may be more abstract than one's own outcomes (Amodio and Frith 2006). In this sense, the present findings support the idea of a dorsal–ventral and posterior–anterior axis (Denny *et al.*, 2012; Suzuki *et al.*, 2012; Koritzky *et al.*, 2013), according to which the more dorsal and anterior mPFC processes more abstract and complex information than the more ventral and posterior mPFC. This in turn is in agreement with the core role of anterior mPFC in social value processing, other-related judgments and mentalizing (Ochsner *et al.*, 2004; Amodio and Frith 2006; Gilbert *et al.*, 2006; Mitchell 2009; Krienen *et al.*, 2010; Fareri *et al.*, 2012). Indeed, the most anterior part of the prefrontal cortex, the frontal pole, may have emerged as a new prefrontal area during primate evolution (Genovesio *et al.*, 2014). Together with the notion that social functions developed to a disproportionate degree in the later stages of primate evolution (Dunbar 1998), it is tempting to speculate that this area might have evolved to serve a preferential role for learning about observed and socially relevant outcomes.

In the domain of causal learning, the blocking effect also occurs in an observational context in which the participant has to learn causal relationships between actions or events and their associated outcomes (Dickinson *et al.*, 1984). We cannot rule out the possibility that the neural results obtained in our study might also generalize to blocking effects in causal learning and may be partly driven by explicit (verbal) reasoning. Previous studies primarily found the lateral PFC to be crucial for causal learning and blocking of causal learning (Fletcher *et al.*, 2001; Turner *et al.*, 2004; Corlett and Fletcher 2012) and verbal reasoning (Costafreda *et al.*, 2006; Tsuchida and Fellows 2013). Extending these studies, we found a more medial region that was specifically involved in blocking in the social over and above the individual domain.

Although the present study focused on blocking in the social domain, we also found individual blocking effects. These were represented primarily in the ventral part of the mPFC and, at less stringent statistical thresholds, also in the striatum and the posterior cingulate (data not shown). These findings replicate our and others' previous reports on individual blocking and learning (Tobler *et al.*, 2006; McDannald *et al.*, 2014; note that in some of the previous studies subjects were thirsty and rewards were drops of liquid, which could have resulted in a more homogeneous and higher value of the reward). By using secondary (monetary) rather than primary rewards to study the neural basis of the blocking effect, we go beyond previous research and show that the vmPFC contributes to individual neural blocking, not only in the context of primary rewards (liquid), but also in that of secondary rewards (money). It should be noted, however, that we did not find significantly stronger activation in the vmPFC for the direct comparison between the individual and social condition. Thus, we cannot conclude that the ventral part of mPFC is specific for processing self-relevant rewards. Indeed, there have also been reports of other-relevant learning processes in the vmPFC (Burke *et al.*, 2010; Suzuki *et al.*, 2012). One possibility worthy of further study is that these ventral regions are engaged when other-relevant learning has a direct benefit (instrumental value) for the observing individual (Burke *et al.*, 2010). In contrast, the observation of others' rewards had comparatively little instrumental value in our study. Thus, it remains to be determined what specific contextual aspects lead to vmPFC contribution during learning in the social domain.

Taken together, our findings substantiate the notion that the same formal learning processes hold and facilitate efficient learning in both the individual and the social domain. Moreover, our data indicate that regions of dorsal mPFC play a preferential role for implementing these processes when rewards are socially relevant.

SUPPLEMENTARY DATA

Supplementary data are available at *SCAN* online.

CONFLICT OF INTEREST

None declared.

REFERENCES

Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Review Neuroscience*, *7*, 268 77.

Becker, G.M., Degroot, M.H., Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*, 226 32.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S. (2008). Associative learning of social value. *Nature*, *456*, 245 9.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624 52.

Burke, C.J., Tobler, P.N., Baddeley, M., Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences United States of America*, *107*, 14431 6.

Corlett, P.R., Fletcher, P.C. (2012). The neurobiology of schizotypy: fronto-striatal prediction error signal correlates with delusion-like beliefs in healthy people. *Neuropsychologia*, *50*, 3612 20.

Costafreda, S.G., Fu, C.H.Y., Lee, L., Everitt, B., Brammer, M.J., David, A.S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: role of the left inferior frontal gyrus. *Human Brain Mapping*, *27*, 799 810.

Denny, B.T., Kober, H., Wager, T.D., Ochsner, K.N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *24*, 1742 52.

Dickinson, A., Shanks, D., Evenden, J. (1984). Judgement of act-outcome contingency: the role of selective attribution. *Quarterly Journal of Experimental Psychology A*, *36*, 29 50.

Dunbar, R.I.M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, *6*, 178 90.

Eippert, F., Gamer, M., Büchel, C. (2012). Neurobiological mechanisms underlying the blocking effect in aversive learning. *Journal of Neuroscience*, *32*, 13164 76.

Fareri, D.S., Niznikiewicz, M.A., Lee, V.K., Delgado, M.R. (2012). Social network modulation of reward-related signals. *Journal of Neuroscience*, *32*, 9045 52.

Fletcher, P.C., Anderson, J.M., Shanks, D.R., et al. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience*, *4*, 1043 8.

Galef, B.G., Durlach, P.J. (1993). Absence of blocking, overshadowing, and latent inhibition in social enhancement of food preferences. *Animal Learning and Behavior*, *21*, 214 20.

Genovesio, A., Wise, S.P., Passingham, R.E. (2014). Prefrontal-parietal function: from foraging to foresight. *Trends in Cognitive Sciences*, *18*, 72 81.

Gilbert, S.J., Spengler, S., Simons, J.S., et al. (2006). Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. *Journal of Cognitive Neuroscience*, *18*, 932 48.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences United States of America*, *10*, 6741 6.

Kamin, L. (1969). Predictability, surprise, attention and conditioning. In: Campbell, B.A., Church, R.M., editors. *Punishment and Aversive Behavior*. New York: Appleton-Century-Crofts, pp. 279–96.

Kerns, J.G., Cohen, J.D., MacDonald III, A.W., Cho, R.Y., Stenger, V.A., Carter, C.S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*, 1023–6.

Koritzky, G., He, Q., Xue, G., Wong, S., Xiao, L., Bechara, A. (2013). Processing of time within the prefrontal cortex: recent time engages posterior areas whereas distant time engages anterior areas. *Neuroimage*, *72*, 280–286.

Krienen, F.M., Tu, P.C., Buckner, R.L. (2010). Clan mentality: evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience*, *30*, 13906–15.

Leathers, M.L., Olson, C.R. (2012). In monkeys making value-based decisions, LIP neurons encode cue salience and not action value. *Science*, *338*, 132–5.

McDannald, M.A., Jones, J.L., Takahashi, Y., Schoenbaum, G. (2014). Learning theory: a driving force in understanding orbitofrontal function. *Neurobiology of Learning and Memory*, *108*, 22–7.

Melchers, K.G., Shanks, D.R., Lachnit, H. (2008). Stimulus coding in human associative learning: flexible representations of parts and wholes. *Behavioural Processes*, *77*, 413–27.

Mitchell, J.P. (2009). Social psychology as a natural kind. *Trends in Cognitive Sciences*, *13*, 246–51.

Ochsner, K.N., Knierim, K., Ludlow, D.H., et al. (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*, *16*, 1746–72.

Prados, J. (2011). Blocking and overshadowing in human geometry learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*, 121–6.

Rescorla, R.A., Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F., editors. *Classical Conditioning II: Current Research and Theory*. New York: Appleton-Century-Crofts, pp. 64–99.

Shenhav, A., Botvinick, M.M., Cohen, J.D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*, 217–40.

Suzuki, S., Harasawa, N., Ueno, K., et al. (2012). Learning to simulate others' decisions. *Neuron*, *74*, 1125–37.

Tobler, P.N., O'Doherty, J.P., Dolan, R.J., Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*, *95*, 301–10.

Tsuchida, A., Fellows, L.K. (2013). Are core component processes of executive function dissociable within the frontal lobes? Evidence from humans with focal prefrontal damage. *Cortex*, *49*, 1790–800.

Turner, D.C., Aitken, M.R.F., Shanks, D.R., et al. (2004). The role of the lateral frontal cortex in causal associative learning: exploring preventative and super-learning. *Cerebral Cortex*, *14*, 872–80.

Waelti, P., Dickinson, A., Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43–48.

Zhu, L., Mathewson, K.E., Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences United States of America*, *109*, 1419–24.

Supplementary data for:

Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex

Azade Seid-Fatemi and Philippe N. Tobler

University of Zurich

Table S1. Trial types and number of trials used in each phase

Phase	Social condition		Individual condition	
	Trial type	**Number of trials**	**Trial type**	**Number of trials**
Pretraining	A_{SOCIAL} +	20	$A_{INDIVIDUAL}$ +	20
	B_{SOCIAL} -	20	$B_{INDIVIDUAL}$ -	20
Compound conditioning	A_{SOCIAL} +	14	$A_{INDIVIDUAL}$ +	14
	B_{SOCIAL} -	14	$B_{INDIVIDUAL}$ -	14
	AX_{SOCIAL} +	24	$AX_{INDIVIDUAL}$ +	24
	BY_{SOCIAL} +	24	$BY_{INDIVIDUAL}$+	24
	CZ_{SOCIAL} -	12	$CZ_{INDIVIDUAL}$ -	12
Test	A_{SOCIAL} +	14	$A_{INDIVIDUAL}$ +	14
	B_{SOCIAL} -	14	$B_{INDIVIDUAL}$ -	14
	AX_{SOCIAL} +	24	$AX_{INDIVIDUAL}$ +	24
	BY_{SOCIAL} +	24	$BY_{INDIVIDUAL}$+	24
	CZ_{SOCIAL} -	12	$CZ_{INDIVIDUAL}$ -	12
	X_{SOCIAL} -	14	$X_{INDIVIDUAL}$ -	14
	Y_{SOCIAL} -	14	$Y_{INDIVIDUAL}$ -	14

Rewarded trials are denoted with (+) and unrewarded trials with (-). A and B refer to pretraining trials, AX and BY to rewarded and CZ to unrewarded compound trials, and X and Y to unrewarded test trials.

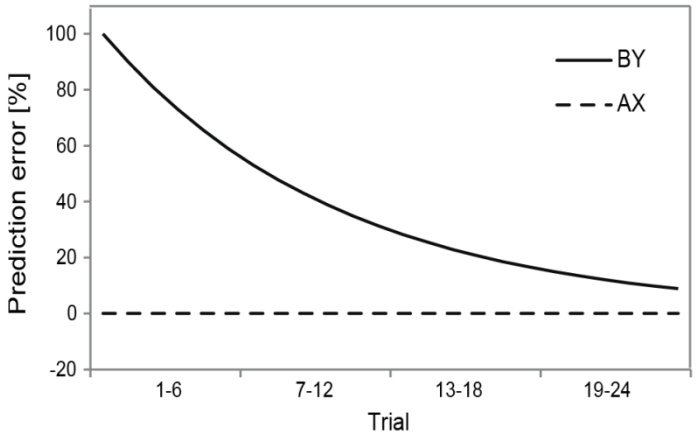

Fig. S1 Theoretically expected prediction errors for BY and AX trials. At the beginning of the compound conditioning phase, BY trials elicit a positive prediction error, which gradually decreases over time as participants learn to associate BY with the reward. In contrast, the prediction error in AX trials should remain constantly close to zero as A trials have already been associated with reward during pretraining. Thus, in AX trials during compound conditioning, there should be no prediction error when rewards are shown. In our fMRI analysis, we captured this effect by contrasting parametric modulators corresponding to decreasing prediction error signals in BY versus AX trials. Thus, by contrasting parametric modulators, we asked which brain region shows a better fit with a decreasing prediction error signal in BY trials than it does in AX trials.

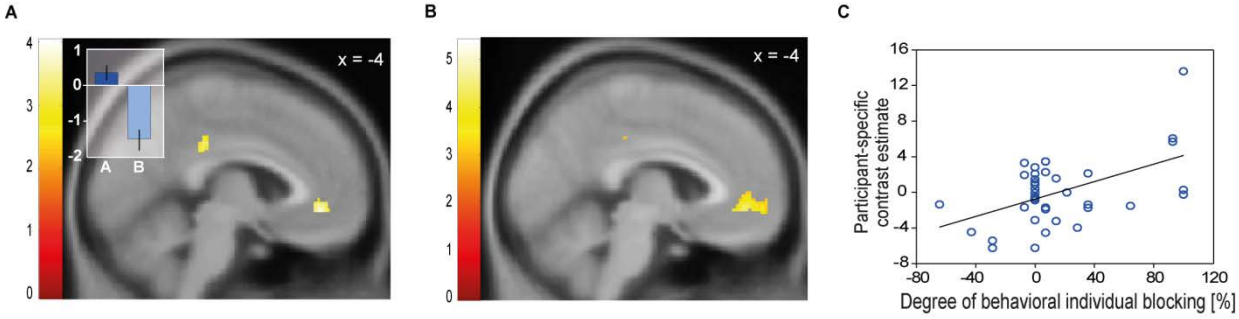

Fig. S2 Activity in vmPFC reflects expectation and blocking for individual rewards. ***A***, vmPFC responses were higher to reward-predicting A stimuli as compared to neutral B stimuli (-4, 40, -6; $p < 0.05$, FWE small-volume corrected). Contrast estimates (inset) show vmPFC responses to A and B stimuli separately. Error bars indicate SEM. ***B*** and ***C***, Differences in activation responses to Y stimuli as compared to X stimuli increased in the vmPFC (-6, 42, -4, $p < 0.05$, FWE small-volume corrected) with degree of behavioral blocking in the individual condition. Blocking was quantified as the difference in reward-expecting responses to non-blocked stimulus Y compared to blocked stimulus X.

B Appendix to Study 2

Social unblocking

Azade Seid-Fatemi and Philippe N. Tobler

Laboratory for Social and Neural Systems Research

Department of Economics

University of Zurich

Abstract

Learning is usually driven by value differences between the outcomes we expect and the outcomes we actually receive. This can be demonstrated with the *blocking effect* in which a novel stimulus is blocked from learning when it is associated with a fully predicted outcome. However, learning can be *unblocked* if a discrepancy between expectation and outcome is introduced. It has been shown that discrepancies in both the value and the identity of the reward can unblock learning but it is unclear whether discrepancies in who actually receives a reward can do so as well. Here, we employed social variants of the unblocking and blocking experiments to assess learning driven by changes in the reward recipient. Participants learned to associate distinct visual cues with value-matched monetary rewards that they or another person received. Subsequently, these cues were compounded with novel visual cues. The reward recipient of the compound trials remained the same as before in the blocking experiment but changed in the unblocking experiment. We found that participants who normally block redundant learning in the standard blocking paradigm show unblocking when the reward recipient changes. Moreover, the degree of recipient-specific unblocking was higher in less empathic and less prosocial participants. Thus, our findings suggest that value-matched learning is sensitive to the reward recipient and that individual differences in unblocking may relate to social traits.

Introduction

While early theories of classical conditioning assumed that the temporal contiguity between stimuli and reward drives learning (Pavlov, 1927), modern learning theories highlight that the discrepancy between the actual and predicted outcome is critical for learning (Rescorla and Wagner, 1972; Mackintosh, 1975; Pearce and Hall, 1980). This is clearly demonstrated by the *blocking effect* (Kamin, 1969). In the blocking paradigm, a stimulus is paired with an outcome whose occurrence is already predicted by another, previously conditioned stimulus. Although the stimulus is closely followed by the outcome, it is redundant and thus "blocked" from learning. This blocking effect occurs because the other, previously learned cue has already been established as a reliable predictor of the outcome and, according to learning theories, there is no discrepancy between the actual and predicted outcome.

Importantly, however, blocking can be prevented if the value (but not the identity) or the identity (but not the value) of the outcomes is changed. In order to show this *unblocking effect*, a variation of the blocking paradigm is used in which the compound of the first and second cue is followed by a different quantity (Holland, 1984) or different identity (Rescorla, 1999) of the outcome than was predicted by the first cue. In contrast to blocking, changing the value or identity of the outcome can result in learning of the second cue due to the discrepancy between expected and obtained value or identity. However, unblocking does not always occur. Indeed, several studies (Bakal et al., 1974; Dickinson and Dearing, 1979; Ganesan and Pearce, 1988) have shown that when the identity but not the value of the outcome is changed the blocking effect is still present. This is the so called *transreinforcer blocking* phenomenon. It suggests that at least during some types of learning only value, as processed by a common motivational system, is relevant for learning, while other features such as identity are abstracted away. Transreinforcer blocking and unblocking have been studied in rats (Burke et al., 2008; McDannald et al., 2011; Steinberg et al., 2013) and in humans during causal learning (Le Pelley et al., 2005). These studies investigated the mechanisms of (un)blocking when learning took place with rewards for the learning individual. However, in real life we often learn in an environment that also provides outcomes for others. It is not clear whether or not social outcome shifts would block learning. More specifically, it is an open question how learning would take place if neither the subjective value, nor the identity of the outcome would change, but its recipient. Switching the recipient of a reward might result in unblocking, or alternatively, as the value and identity of the outcome are the same, lead to transrecipient blocking. To address this

question, we developed a transrecipient variant of the (un)blocking paradigm in which we changed the recipient of the monetary rewards while keeping their subjective value constant across individual and social conditions.

How we learn about social information might depend also on factors that are not related to the outcome itself, but to the degree to which we experience similar feelings from social and individual outcomes. Thus, social learning might depend on our capacity to experience empathy, a relationship that has been demonstrated in recent studies (Fukushima and Hiraki, 2009; Rak et al., 2013; Gossen et al., 2014). We therefore assessed participants' empathy and hypothesized that people low in empathy will be more sensitive to recipient changes, as they perceive the consequences of others' outcomes similarly to the consequences of their own outcomes.

Materials and Methods

Participants and procedure

Thirty-four participants (15 female; aged 20.9 ± 0.4 years; range: 18-28) took part in this study. None of the participants had prior histories of neurological or psychiatric disorders and all had normal or corrected-to-normal vision. Written informed consent was obtained from all participants, and the study was approved by the Research Ethics Committee of the Canton of Zurich.

All participants performed a blocking (Seid-Fatemi and Tobler, 2014) and a transrecipient (un)blocking experiment in two separate sessions on different days. The delay between the two sessions varied between two and three weeks and the order of the two experiments was counterbalanced between participants.

Experimental Design

Stimuli and Outcomes

In both experiments we used individual and social monetary rewards. Individual rewards were received by the participant, social rewards by another person. Two female volunteers served as the other person in social outcome conditions. To ensure that the consequences of the social rewards were as real as those of the individual rewards we used existing persons who received money at the end of the experiment. Using two rather than only one volunteer served to make

the social conditions more engaging (i.e. more varied and less monotonous) and thereby prevent adaptation. The participants never met the two volunteers face-to-face, but read a brief description of them before the experiment began, which included their initials and information about their gender and age.

Before the experiment, we determined the amount of the individual reward according to the social preferences of each subject. We did this to ensure that the rewards in social and individual conditions had the same subjective value. To achieve this, we used a variant of the Becker–DeGroot–Marschak method (BDM; Becker et al., 1964). Specifically, before the experiment, we asked subjects to indicate the amount of money (between CHF 1 and 100) that, if delivered to them, was as valuable as delivering CHF 60 to the other person. The amount of CHF 60 was chosen based on pilot studies with a separate set of subjects showing that CHF 60 yielded affordable individual equivalence amounts (CHF 45.80 ± 1.90). The bid was then compared to a random number between 1 and 100 generated by the computer. If the number was greater than or equal to the subject's bid, they received the indicated amount of money. If the number was lower than the bid, they received nothing and the other person received CHF 60. Thus, the procedure provided an incentive-compatible way of obtaining individual reward amounts that corresponded to the value of social reward amounts. The outcome of the procedure had no influence on the payout or the number of rewards the participants gained from the actual experiment. The bid was obtained before the experiment and used to set the individual reward amount in the experiment such that it had the same value as the social reward amount, given the subject's social preferences.

During the actual experiment the delivery of social or individual outcomes followed the presentation of visual stimuli (Figure 1A). All of the stimuli used were abstract colored shapes presented on a white background and were similar to those used in previous blocking experiments (Waelti et al., 2001; Tobler et al., 2006). Each trial had either a social or an individual outcome, never both. Different stimuli were used for each of the experiments (blocking and transrecipient (un)blocking) and conditions (social and individual).

Blocking experiment

Blocking was tested with the standard within-subject paradigm (Waelti et al., 2001; Tobler et al., 2006) and comprised three consecutive phases. In each of these phases, participants were presented with visual stimuli that were associated with social or individual outcomes (Figure 1A; and see Seid-Fatemi and Tobler, 2014).

In the first (pretraining) phase, the A (experimental) stimuli (A_{SOCIAL} and $A_{INDIVIDUAL}$) were paired with a social or individual reward. In contrast, the B (control) stimuli (B_{SOCIAL} and $B_{INDIVIDUAL}$) were not paired with a reward. Stimuli were presented 20 times each and the identities of the stimuli were counterbalanced across participants. Each trial started with a 4-s intertrial interval (ITI) that varied from 2 to 6 s (Figure 1B). Stimuli were presented for 1.5 s at random either to the left or the right of the fixation cross. The outcome was presented concurrently with the stimulus for another 1.5 s. During the presentation of any given stimulus, the participants were to perform a specific key press corresponding to the recipient and to the outcome that would follow the stimulus. In particular, upon each stimulus presentation, participants had to indicate whether they expected reward for self, no reward for self, reward for others, or no reward for others by pressing a key with the index or middle finger of their left or right hand. Thus, there was an individual and social reward key and an individual and social no-reward key and participants were asked to press one of these keys in each trial. This allowed us to measure recipient- as well as outcome-specific learning. Condition-to-hand (individual or social) and key-to-reward (reward or no reward) assignments were counterbalanced across participants. Trials in which the participant failed to respond or responded too late were repeated later. Visual stimuli as well as response recordings were controlled using Cogent 2000 (Wellcome Department of Imaging Neuroscience, London, UK) as implemented in Matlab.

The remaining parts of the blocking experiment were conducted in the fMRI scanner (data not shown here) and therefore visual stimuli were presented on a display that participants viewed via a mirror fitted to the top of the head coil. In the compound phase, A stimuli were presented together with X stimuli (X_{SOCIAL} and $X_{INDIVIDUAL}$), forming rewarded compounds. As a control, B stimuli were presented together with Y stimuli (Y_{SOCIAL} and $Y_{INDIVIDUAL}$) and also followed by a reward. In AX trials, the upcoming reward was predicted by the A stimuli and therefore should be blocked from learning. In contrast, in early BY trials, the B stimuli did not predict a reward and the outcome of these trials was therefore more valuable than predicted. Accordingly, Y stimuli, but not X stimuli, should be learned as reward-predicting stimuli. AX and BY trials were presented in 24 trials per condition and intermixed with 14 A and B trials per condition, which served to maintain the previously learned associations. We also included control compound trials (CZ trials) that were unrewarded (12 trials each for the social and individual condition). We used this trial type to prevent compound trials in general from being associated with reward and to ensure that participants paid attention to each of the individual

stimuli that constituted a compound rather than automatically associating the co-occurrence of any two stimuli with reward.

In a third phase, X and Y stimuli were presented alone in unrewarded test trials. Under the assumption that previous learning blocks subsequent learning, the X stimuli should have been blocked from being associated with social or individual reward, while the Y stimuli should have been associated with reward. Y and X trials were presented in 14 trials each and randomly intermixed with A and B trials (14 trials), AX and BY trials (24 trials), and control compound trials (12 trials), again, to maintain previously learned associations. As before, A, AX, and BY trials were followed by reward in order to maintain the previously learned associations.

Transrecipient-(un)blocking experiment

Like the blocking experiment, the transrecipient (un)blocking experiment comprised three phases. The first, pretraining phase was identical to the one in the blocking experiment. However, the second, compound conditioning phase differed from the standard blocking experiment. Here, the identity of the reward recipient was changed when the pretrained stimuli A were presented in compound with stimuli X. In the social (un)blocking condition an $A_{INDIVIDUAL}$ stimulus was pretrained with a monetary outcome for the participant, but then resulted in a reward (of equal value) for another person when presented with a second stimulus during compound conditioning (AX_{SOCIAL}). Conversely, in the individual (un)blocking condition pretrained rewards were received by another person (A_{SOCIAL}) and the recipient switched to the participant in the compound phase ($AX_{INDIVIDUAL}$) (Figure 1A).

In the control condition the recipient of the outcome was changed in a similar fashion. After participants had learned that control stimulus B predicted no reward for one recipient (e.g. individual), the presentation of stimuli B and Y in compound was followed by reward for the other recipient (e.g. social). As in the blocking experiment, in the third phase, X and Y stimuli were presented alone in unrewarded test trials. Although we did not further analyze the control conditions (B, BY and Y) in the (un)blocking experiment, we kept them in the experiment in order to make the (un)blocking experiment as similar as possible to the blocking experiment. Thus, other than the change of reward recipient in the compound conditioning phase the transrecipient (un)blocking was identical to the blocking experiment with respect to trial structure and trial number.

Payment

Participants were instructed that, at the end of the experiment, a portion of the rewards accumulated in correctly predicted trials would be paid out to them and the other two individuals, respectively. To ensure that everyone received approximately the same amount irrespective of their bid in the BDM, we adjusted the percentage for each participant individually. To keep them engaged throughout the experiment, in each trial in which participants failed to respond or responded too slowly, CHF 1 was deducted from their final monetary payment and the three participants with the highest number of correct responses received an additional payment (CHF 20).

Assessment of Empathy and Prosociality

After completion of the experiment we assessed participants' trait empathy with a questionnaire (Interpersonal Reactivity Index, IRI; Davis, 1983) which comprises 28 items. This multidimensional approach measures empathy with different subscales that assess perspective taking, fantasy, empathic concern and personal distress. Additionally, participants rated their prosociality on a 16-item scale (Caprara et al., 2005) that assessed their propensity to help, share, console, support, and cooperate with others.

Data analysis

Social and individual reward expectations were defined as the percentage of the social and individual reward key pressed, respectively, and were evaluated using paired t-tests and two-way repeated-measures ANOVAs. Keypress data were arcsine-square root transformed (i.e. were variance-stabilized) to permit the application of parametric statistical tests. The degree of participant-specific behavioral blocking was calculated as the difference between recipient-specific reward key presses for Y stimuli and those for X stimuli. Comparing the responses to Y with those to X stimuli is the standard approach to determining whether or not and to what degree blocking has taken place (Tobler et al., 2006; Eippert et al., 2012). The larger the difference, the stronger the blocking effect. Participants were classified as blockers if this difference was positive.

To detect and quantify transrecipient (un)blocking we followed an approach used previously (McDannald et al., 2011) to detect transreinforcer (un)blocking. Specifically, responses to the X stimulus in the (un)blocking experiment were compared to responses to the

X stimulus in the blocking experiment. For each condition (social and individual) we compared the X stimuli with the same recipient. That is, in the individual condition we compared the X stimuli that in both experiments were followed by individual reward in AX compounds. While in the blocking experiment the A stimulus also predicted individual reward in pretraining trials, it predicted social reward in pretraining trials of the transrecipient (un)blocking experiment. In the social condition we used the same approach and compared the X stimuli that in both experiments were followed by social reward in AX compounds. Increased reward key responses to the X stimulus in the transrecipient (un)blocking experiment as compared to the standard blocking experiment indicate an unblocking effect, while similar levels of responses would indicate transrecipient blocking. Thus, the larger the difference, the stronger the unblocking effect.

Results

We used a within-subject design and two experiments, performed on different days, to test blocking and transrecipient (un)blocking effects with social and individual outcomes (Figures 1A and 1B, see Methods). During the pre-training phase of both experiments, participants learned to associate one stimulus (A) with reward. In the blocking experiment the pre-trained association between A and reward was expected to block learning to X in the compound phase, but it was an open question whether blocking would also occur in the transrecipient (un)blocking experiment, in which during compound conditioning the recipient of the reward was changed from individual to social or vice versa. Lower responding would be compatible with a transreinforcer blocking effect. Alternatively, changing the recipient of the reward during compound conditioning might not block responding to X, compatible with an unblocking effect.

As reported previously (Seid-Fatemi and Tobler, 2014), we found a blocking effect, measured as increased reward key pressing to Y than X stimuli in both the social ($t_{(33)} = 2.98$, $p < 0.01$) and the individual ($t_{(33)} = 2.63$, $p < 0.05$) condition of the standard blocking experiment. The average amount of blocking was similar for the social and individual condition ($t_{(33)} = 0.11$, $p = 0.92$). However, for both conditions the degree of blocking varied across participants. Out of the 34 subjects, 17 (19) showed responses compatible with an individual (social) blocking effect, that is more key presses for individual (social) Y stimuli than for individual (social) X stimuli. Moreover, across participants, the degree of blocking in the

social condition was correlated with the degree of blocking in the individual condition ($R^2 = 0.53$, $p < 0.001$), suggesting a relation between social and individual blocking.

Next, we investigated responses in the transrecipient (un)blocking experiment. Before testing for a transrecipient (un)blocking effect we confirmed that participants had learned the (previous) stimulus-outcome associations during the pretraining phase: Participants predicted recipient-specific reward outcomes when presented with reward-predicting A stimuli, but not when presented with B (control) stimuli. This resulted in a significantly higher number of reward key presses for A vs. B stimuli, for both social and individual conditions (social: $t_{(33)} = 12.26$; individual: $t_{(33)} = 24.69$, both $p < 0.001$), indicating that the participants learned to discriminate in an outcome- and recipient-specific manner between stimuli predicting reward and stimuli predicting no reward.

To measure whether transrecipient or unblocking effects occurred, we compared responses to the X stimulus in the transrecipient (un)blocking experiment to the X stimulus in the blocking experiment. When including all subjects irrespective of their behavior in the blocking experiment we observed that stimulus X is unblocked when individual changes to a social reward outcome, as participants showed increased reward-expecting button presses for stimulus X in the (un)blocking experiment compared to the blocked stimulus X in the standard blocking experiment (Figure 2A; $t_{(33)} = 2.11$, $p < 0.05$). By contrast, for the individual condition our data indicate a transrecipient blocking effect as responses to stimulus X in the unblocking experiment do not differ significantly from the responses to stimulus X in the blocking experiment (Figure 2A; $t_{(33)} = 0.73$, $p = 0.47$). The suggested experiment-dependent difference in the degree of blocking reached trend-level significance ($F_{(1,33)} = 3.62$; $p = 0.07$) when assessed with a condition by experiment interaction.

As we find individual differences with respect to the occurrence of the blocking effect we next tested for a transrecipient (un)blocking effect depending on behavior in the blocking experiment. In particular, we asked whether subjects who showed blocking effects in the standard blocking experiment would also show reduced reward key responses to the X stimulus in the transrecipient (un)blocking experiment, or whether they would show increased responses to the X stimulus, compatible with an unblocking effect. We observed that stimulus X is unblocked when social valuable outcomes change to individual rewards, as participants showed increased reward-expecting button presses for stimulus X in the unblocking experiment compared to the blocked stimulus X in the standard blocking experiment (Figure 2B; $t_{(16)} = 2.13$, $p < 0.05$). For the social condition we found similar unblocking effects when individual changed

to social rewards, albeit at a trend level (Figure 2B; $t_{(18)} = 1.92$, $p < 0.07$). This suggests that participants who show blocking in the standard paradigm demonstrate unblocking when the reward recipient changes and this relation was confirmed in a correlation analysis. Indeed we found a positive correlation between the degree of blocking and the degree of unblocking for both the individual ($R^2 = 0.56$, $p < 0.001$; when including all participants: $R^2 = 0.46$, $p < 0.001$) and the social ($R^2 = 0.35$, $p < 0.005$; when including all participants: $R^2 = 0.31$, $p < 0.005$) condition. Moreover, the degree of unblocking in the individual condition was correlated with the degree of unblocking in the social condition ($R^2 = 0.81$, $p < 0.001$; when including all participants: $R^2 = 0.69$, $p < 0.001$). Taken together, our data indicate that participants who show standard blocking, exhibit an unblocking effect when the recipient of the reward is changed and this effect was related between the individual and social conditions.

Finally, we asked whether individual differences in empathy or prosociality can predict the degree of unblocking. We found that the empathic concern scale correlated negatively with the degree of unblocking in both the individual (Figure 3A; $R^2 = 0.23$; $p < 0.01$) and social condition (Figure 3B; $R^2 = 0.17$; $p < 0.05$), indicating that participants with less empathic concern showed more unblocking. Similarly, although only at trend level, we found a negative correlation between prosociality and the degree of unblocking in both the individual ($R^2 = 0.10$; $p = 0.07$) and social condition ($R^2 = 0.10$; $p = 0.07$). Moreover, for the individual but not the social condition a negative correlation was found with the personal distress scale of the IRI ($R^2 = 0.12$; $p < 0.05$). No significant correlations were found for the other subscales (all $p > 0.12$).

Discussion

The present study investigated the unblocking effect in a social context by using a novel paradigm in which the recipient of the monetary outcome was changed without accompanying changes in subjective value or reward identity (money). We found that participants who normally block redundant learning in the standard blocking paradigm show unblocking when the reward recipient is changed. Thus, although social and individual learning follow similar learning principles and both show correlated blocking effects, participants are sensitive to who receives reward and track changes by showing unblocking.

When investigating the behavior of all participants, we observed unblocking for the social but not the individual condition, suggesting that unblocking occurs preferentially when individual rewards change to socially relevant rewards. Our finding seems to contradict previous work showing a bias toward a reliance on individual information (Eriksson and Strimling, 2009; Morgan et al., 2012). This bias would predict an opposite pattern, as participants should rely more on individual cues and outcomes when the outcome changes from individual to social and vice versa. Thus, if they were more sensitive to learning from individual information, participants should have shown unblocking when reward receipt changed from social to individual and blocking when receipt changed in the reverse direction. It is conceivable that the degree of reliance on individual information depends on the degree of competition between individual and social information but future research is needed to elucidate this possibility.

Interestingly, lower scores in empathy and prosociality were associated with greater unblocking effects, suggesting that less empathic and prosocial individuals are more sensitive to who actually receives rewards. In other words, more empathic concern leads individuals to treat others' rewards more similarly to their own and higher selfishness leads to increased differentiation between rewards received by self and other. This interpretation is supported by previous research that has linked empathy to altruism (Batson et al., 1991, 2003) and sharing behaviors (Edele et al., 2013). Our empathy finding should be interpreted with caution, however, since we only find a correlation with the empathic concern subscales of the IRI. In particular, our initial hypothesis would predict that also the perspective taking subscale should relate to unblocking effects. Empathy is currently thought of as a multidimensional concept, involving a cognitive and an affective component (Davis, 1983; Decety and Lamm, 2006). While perspective taking represents the cognitive aspect, empathic concern reflects the affective aspect of empathy. Thus, our data seem to indicate that not understanding others' perspective but the emotional reaction to others' needs and welfare accounts for differences in the degree of unblocking in a social context. Future studies are needed to investigate this relation in greater detail. Taken together, our findings suggest that reward recipient unblocking occurs during social learning and may be related to individual variability in social traits.

References

Bakal CW, Johnson RD, Rescorla RA (1974) The effect of change in US quality on the blocking effect. Pav J Biol Sci 9:97–103.

Batson CD, Batson JG, Slingsby JK, Harrell KL, Peekna HM, Todd RM (1991) Empathic joy and the empathy-altruism hypothesis. J Pers Soc Psychol 61:413–426.

Batson CD, Lishner DA, Carpenter A, Dulin L, Harjusola-Webb S, Stocks EL, Gale S, Hassan O, Sampat B (2003) "... As you would have them do unto you": Does imagining yourself in the other's place stimulate moral action? Pers Soc Psychol Bull 29:1190–1201.

Becker GM, Degroot MH, Marschak J (1964) Measuring utility by a single-response sequential method. Syst Res 9:226–232.

Burke KA, Franz TM, Miller DN, Schoenbaum G (2008) The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. Nature 454:340–344.

Caprara GV, Steca P, Zelli A, Capanna C (2005) A new scale for measuring adults' prosocialness. Eur J Psychol Assess 21:77–89.

Davis MH (1983) Measuring individual differences in empathy: Evidence for a multidimensional approach. J Pers Soc Psychol 44:113–126.

Decety J, Lamm C (2006) Human empathy through the lens of social neuroscience. Sci World J 6:1146–1163.

Dickinson A, Dearing MF (1979) Appetitive-aversive interactions and inhibitory processes. In: Mechanism of learning and motivation (Dickinson A, Boakes RA, eds), pp 203–231. Hillsdale, NJ: Erlbaum.

Edele A, Dziobek I, Keller M (2013) Explaining altruistic sharing in the dictator game: The role of affective empathy, cognitive empathy, and justice sensitivity. Learn Individ Differ 24:96–102.

Eippert F, Gamer M, Büchel C (2012) Neurobiological mechanisms underlying the blocking effect in aversive learning. J Neurosci 32:13164–13176.

Eriksson K, Strimling P (2009) Biases for acquiring information individually rather than socially. J Evol Psychol 7:309–329.

Fukushima H, Hiraki K (2009) Whose loss is it? Human electrophysiological correlates of non-self reward processing. Soc Neurosci 4:261–275.

Ganesan R, Pearce JM (1988) Effect of changing the unconditioned stimulus on appetitive blocking. J Exp Psychol Anim Behav Process 14:280–291.

Gossen A, Groppe SE, Winkler L, Kohls G, Herrington J, Schultz RT, Grunder G, Spreckelmeyer KN (2014) Neural evidence for an association between social proficiency and sensitivity to social reward. Soc Cogn Affect Neurosci 9:661–670.

Holland PC (1984) Unblocking in Pavlovian appetitive conditioning. J Exp Psychol Anim Behav Process 10:476–497.

Kamin L (1969) Predictability, surprise, attention, and conditioning. In: Punishment and aversive behavior (Campbell BA, Church RM, eds), pp 279–296. New York: Appleton-Century-Crofts.

Le Pelley ME, Oakeshott SM, McLaren IPL (2005) Blocking and unblocking in human causal learning. J Exp Psychol Anim Behav Process 31:56–70.

Mackintosh NJ (1975) A theory of attention: Variations in the associability of stimuli with reinforcement. Psychol Rev 82:276–298.

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G (2011) Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci 31:2700–2705.

Morgan TJH, Rendell LE, Ehn M, Hoppitt W, Laland KN (2012) The evolutionary basis of human social learning. Philos Trans R Soc Lond B Biol Sci 279:653–662.

Pavlov IP (1927) Conditioned reflexes. London: Oxford UP.

Pearce JM, Hall G (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87:532–552.

Rak N, Bellebaum C, Thoma P (2013) Empathy and feedback processing in active and observational learning. Cogn Affect Behav Neurosci 13:869–884.

Rescorla RA (1999) Learning about qualitatively different outcomes during a blocking procedure. Anim Learn Behav 27:140–151.

Rescorla R, Wagner A (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II: Current Research and Theory (Black A, Prokasy W, eds), pp 64–99. New York: Appleton-Century-Crofts.

Seid-Fatemi A, Tobler PN (2014) Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex. Soc Cogn Affect Neurosci. Advance online publication. doi: 10.1093/scan/nsu130.

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH (2013) A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16:966–973.

Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2006) Human neural learning depends on reward prediction errors in the blocking paradigm. J Neurophysiol 95:301–310.

Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412:43–48.

Figure legends

Fig. 1 Experimental design. ***A***, Three phases of standard blocking (left) and transrecipeint (un)blocking (right) paradigm with monetary rewards. The individual conditions are shown as an example. During pretraining, participants learned to associate stimuli with the presence (+) or absence (-) of individual monetary outcomes. Reward-predicting A stimuli were followed by a monetary reward for self, but not neutral B stimuli. During compound conditioning, X and Y stimuli appeared together with A and B stimuli in rewarded compounds. Importantly, while in the standard blocking paradigm the compound stimuli predicted reward for the same recipient

as during previous pretraining (here individual), in the transrecipient (un)blocking paradigm compound stimuli predicted reward for the other recipient (here social). During the test phase, X and Y stimuli were presented alone and remained unrewarded. Rewarded trials are denoted with (+) and unrewarded trials with (-). ***B***, Example of pretraining trials. Abstract visual stimuli were presented in random order, either to the left or the right of the fixation cross. Upon presentation of a stimulus, the participants were to perform a specific key press corresponding to the recipient (self or other) and to the outcome (reward or no reward) that would follow the stimulus. The outcome was shown together with the stimulus for another 1.5 s. The ITI varied between 2 and 6 s.

Fig. 2 Behavioral results. ***A***, Reward-expecting responses (quantified as percentage of key presses) of all subjects. In the social condition participants showed an increase in reward-expecting responses to stimulus X in the transrecipient (un)blocking experiment compared to the blocked stimulus X in the standard blocking experiment, suggesting an unblocking effect when rewards change from individual to social. In the individual condition there was no significant difference between reward-expecting responses to the X stimuli when comparing the standard blocking with the transrecipient (un)blocking experiment. ***B***, Reward-expecting responses of subjects showing blocking in the standard blocking paradigm. In both conditions participants showed increased reward-expecting button presses for stimulus X in the unblocking experiment compared to the one in the standard blocking experiment. Error bars indicate SEMs.

Fig. 3 Negative correlation between the empathic concern score and the degree of unblocking. ***A*** and ***B***, Across all subjects the degree of individual (A) and social (B) unblocking was higher with lower empathy in the empathic concern scale. Unblocking was defined as the difference between reward-expecting responses to stimulus X in the transrecipient (un)blocking experiment compared to stimulus X in the standard blocking experiment.

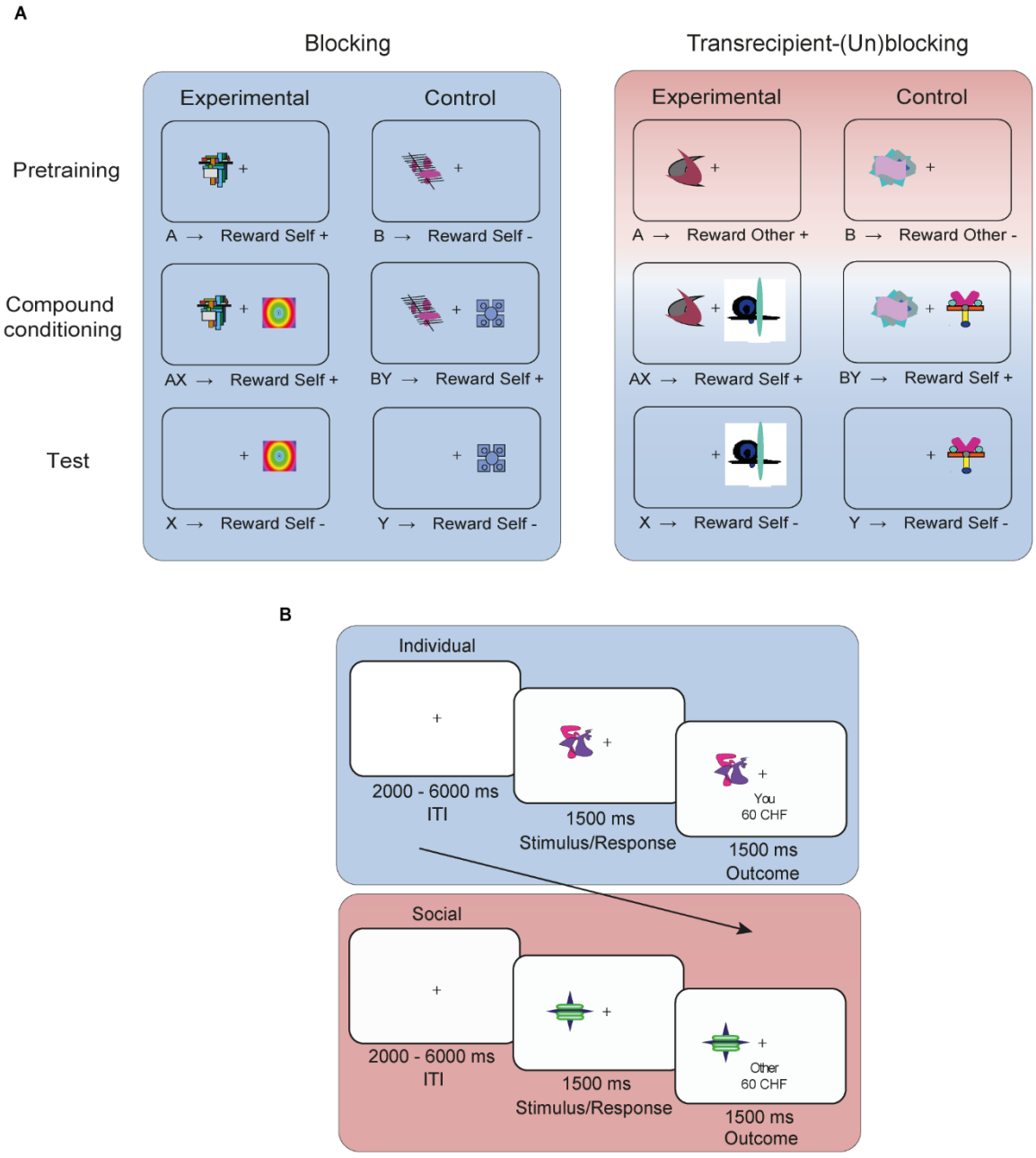
A
Blocking
Transrecipient-(Un)blocking
Experimental
Control
Experimental
Control
Pretraining
Compound conditioning
Test
A → Reward Self +
B → Reward Self -
A → Reward Other +
B → Reward Other -
AX → Reward Self +
BY → Reward Self +
AX → Reward Self +
BY → Reward Self +
X → Reward Self -
Y → Reward Self -
X → Reward Self -
Y → Reward Self -
B
Individual
2000 - 6000 ms
ITI
1500 ms
Stimulus/Response
You
60 CHF
1500 ms
Outcome
Social
2000 - 6000 ms
ITI
1500 ms
Stimulus/Response
Other
60 CHF
1500 ms
Outcome

Figure 1

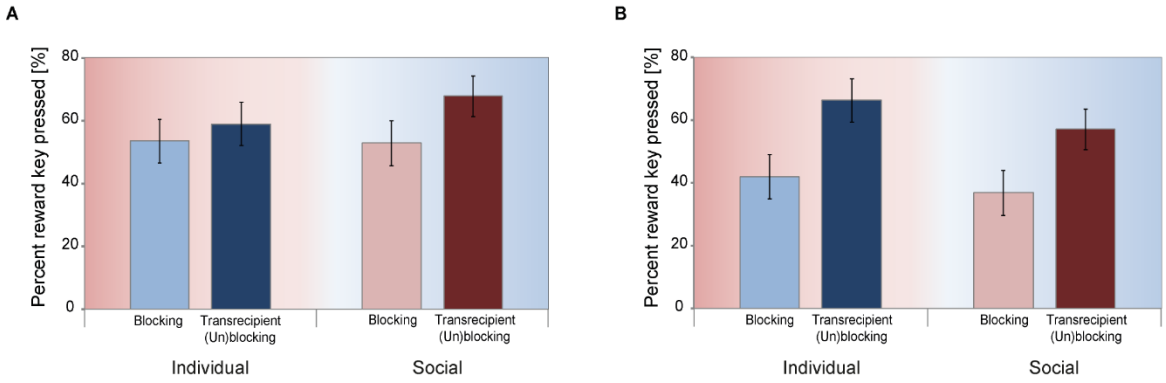

Figure 2

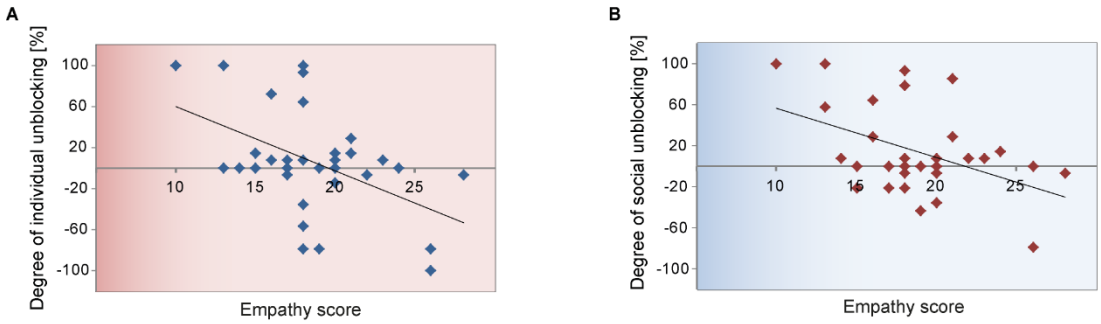

Figure 3

C Appendix to Study 3

Prefrontal connectivity predicts individual differences in honesty

Azade Seid-Fatemi[1], Felix Heise[2], Carmen Tanner[3], Rajna Gibson[3], Alexander F. Wagner[3], Philippe. N. Tobler[1]

[1] Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich,

[2]Department of Psychology, University of Hildesheim,

[3]Department of Banking and Finance, University of Zurich

Abstract

Individuals differ profoundly when they decide whether to tell the truth or to be dishonest. These individual differences can be viewed as personality traits that reflect how individuals will behave when truthfulness is costly. “Protected values” represent such a personality trait that denotes individual propensity to adhere to a principle that shields moral decisions against consideration of (economic) consequences. Recent neuroimaging studies have identified a network of brain regions that are involved in decisions concerning honesty. However, it is not clear how brain regions dynamically communicate with each other in response to variations in economic costs of truthfulness and whether these communications depend on individual personality differences. To investigate these processes we used functional magnetic resonance imaging and measured brain activations while participants decided whether to tell the truth. We found that participants showed stronger parametric cost coding in the dorsolateral and dorsomedial prefrontal cortex the more honest they were. Importantly, in participants with high protected values functional coupling between these regions and the inferior frontal gyrus was tighter in high cost conditions compared to low cost conditions. This finding suggests that moral personality can affect the deployment of control mechanisms. Furthermore, personality-dependent connectivity between the dorsolateral prefrontal cortex and the inferior frontal gyrus was specifically found during moral decisions. Our data provide novel insights into how prefrontal connectivity predicts individual variability in honesty and stress the importance of investigating functional connectivity in studies related to moral decisions about honesty.

Introduction

Decisions concerning honesty are one of the most common moral decisions individuals make, not only in everyday situations but also in contexts that have great economic and political impact. As with decisions in general, decisions about honesty often require weighing the benefits of an option against its costs. Thus, while individuals value the mere act of telling the truth, they often also care about the consequences of their actions. Importantly, individuals differ in the extent to which they consider the consequences of honesty, which leads to individual differences in the tendency to tell the truth. Accordingly, individual differences in honesty are more pronounced when the cost of telling the truth is higher. For most individuals the benefits of telling the truth will outweigh the costs as long as these costs are low. However, when the costs increase, some individuals will no longer be willing to incur them, while others will stand by their moral principles and react less to the costs of doing so.

Recent neuroimaging studies identified a network of brain regions that are involved when subjects make honest or dishonest decisions (for review see Sip et al., 2008; Abe, 2011). In spite of the variety of different experimental paradigms used, such as instructed lying (Spence et al., 2001; Ganis et al., 2003; Langleben et al., 2005; Abe et al., 2006) or spontaneous lying (Baumgartner et al., 2009; Greene and Paxton, 2009; Abe and Greene, 2014), the results consistently revealed contributions of the dorsolateral prefrontal cortex (dlPFC), the dorsomedial prefrontal cortex (dmPFC) as well as the ventrolateral prefrontal cortex (vlPFC) and the inferior frontal gyrus (IFG) (Sip et al., 2008; Abe, 2009). Most of the studies found these regions to be activated when individuals engaged in dishonest behavior as compared with non-deceptive control behavior (Sip et al., 2008; Baumgartner et al., 2009; Abe, 2011). Others observed these regions to be activated when participants voluntarily told the truth as compared to a control condition in which they had no other choice than telling the truth (Sip et al., 2010), particularly in those participants who lied in most trials (Greene and Paxton, 2009).

So far, studies looked primarily at the neural activations underlying honesty in group analyses, without directly considering individual differences in the propensity to tell the truth. To date, only one study has explored the neural mechanisms underlying individual differences in spontaneous dishonest behavior. Abe and Greene (2014) found that individuals with stronger nucleus accumbens responses to anticipated reward show higher levels of dishonest behavior and exhibit greater dlPFC activation when refraining from dishonest behavior. While this study

provides important insights in how the lure of economic incentives influences truthful decisions it is an open question how protection against that lure by strong moral principles is achieved by the brain. Similar to a personality dimension, individuals differ in the degree to which principles guide their moral decisions and these differences are captured by the notion of "protected values" (Baron and Spranca, 1997). More specifically, it has been shown that protected values shield honest decisions against economic considerations, specifically when the costs of truthfulness are high (Gibson et al., 2013). While protected values have been studied on the behavioral level and captured with formal economic models, it is not clear how individual differences in protected values and their impact on honesty are expressed in the brain.

To address these questions we used functional magnetic resonance imaging (fMRI) to measure brain activations while participants made decisions concerning truthfulness. First, by varying the cost of truthfulness we were able to identify brain regions that capture individual differences in the propensity to tell the truth as a function of variations in the economic costs of doing so. Second, by measuring participants' protected values we were able to characterize the neural mechanisms by which protected values impact behavior. Given that truthfulness decisions depend on several variables, such as personality trait and situational costs, it is likely that they are not implemented by a single brain region, but by dynamic changes in the connectivity among different brain regions. Specifically, we hypothesized that individuals with high protected values exhibit a specific prefrontal connectivity pattern that helps them to protect their moral values against the lure of economic incentives when the costs of truthfulness are high. Third, our experimental design also allowed us to address a central question in the study of morality, namely the extent to which moral and non-moral decision making relies on similar or different neural mechanisms. While it has been argued that moral decision making relies on common ("domain-general") rather than particular ("domain-specific") neural mechanisms (Greene and Haidt, 2002; Tobler et al., 2008; Shenhav and Greene, 2010), there have also been proposals that morality forms a domain of its own (Hauser, 2006). If so, we hypothesized that particularly individuals high in protected values should tap into this specific domain and prefrontal connectivity should express it.

Materials and Methods

Participants

Thirty-two subjects (14 female; aged 22.0 ± 0.5 years; range: 18-28) took part in the experiment. None of them had prior histories of neurological or psychiatric disorders and all had normal or corrected-to-normal vision. The study was approved by the Research Ethics Committee of the Canton of Zurich, and all subjects provided informed consent.

Experimental Design

Task instructions and payment

Participants were first asked to read the instructions and answer control questions in order to ensure that they understood the tasks and the rules of the experiment. Before entering the scanner the participants completed a series of practice trials. In the scanner participants performed the truthtelling task and two control tasks (the valuation and the effort task) that were matched to the truthtelling task in terms of potential earnings and costs but did not contain a moral component. Throughout the experiment, participants were placed in the situation of an imaginary CEO who had to make various incentivized decisions. Participants received a fixed basic payment of CHF 25 and a variable additional payment that corresponded to the summed earnings from five randomly selected trials per task and was determined with the actual decisions made in these trials. The average variable payment was CHF 69.60.

In each trial of the truthtelling task the participants had to announce the company's earnings per share for the previous quarter. The participants were informed that the variable component of their payment as a CEO would be higher if they announced higher earnings. They were also told that the market and the shareholders anticipated earnings of 35 cents per share, but that the true earnings were 31 cents per share. Accordingly, participants could either announce 31 cents per share at an economic cost to them (see below) or announce earnings of 35 cents per share while remaining within legal accounting limits. Thus, participants had to choose whether to honestly announce true earnings or to lie and falsely report higher earnings. Importantly, false reports led to higher, whereas truthful reports led to lower actual payoffs for the participants, corresponding to a trade-off between moral and economic incentives. This setup parallels a real-life conflict that CEOs face, as their variable compensation is often tied to stock price performance, which

in turn depends on the earnings they announce. Thus, despite knowing that it is ethically wrong the CEO has economic incentives to behave dishonestly and boost earnings in order to increase his or her payout. Importantly, the CEOs' preferences for telling the truth should decrease as the costs of telling the truth increase.

The bonus to be gained from giving truthful reports varied across trials between CHF 100,000 and CHF 500,000 (see below). The bonus to be gained from giving false reports was fixed to CHF 500,000. These amounts reflect the substantial monetary consequences that earnings management can have for CEOs of public corporations. Participants were informed that for their experimental payoff the CEO bonus would be converted into real money at a rate of CHF 100,000 = CHF 1.

In the two control tasks, too, the decisions the participants made as a CEO affected stock value and therefore their compensation. In the valuation task participants chose between two projects to invest in. The two projects were similar in their properties but differed in their profit and accordingly the CEO compensation. The high value project (labelled project "XIR" in the task) led to a higher actual payment for the CEO (CHF 500,000) than the low value project (labelled "ZEM"; CHF 100,000 – CHF 500,000).

In the effort task participants had to choose how much work to invest as a CEO. The participants were told that the more they worked the more the company would profit and the more they would earn as a CEO. Specifically, in each trial participants decided how many math problems to solve after the experiment outside of the scanner. They chose between solving only one problem or five problems, which would take five times longer than solving one problem. Yet, the five problems option led to a higher actual payment for the CEO (CHF 500,000) than the one problem option (CHF 100,000 – CHF 500,000), resulting in a trade-off between time and economic incentives. Please note that in all tasks and trials participants decided between a variable (1-5 CHF) and a fixed payoff option (5 CHF).

Trial structure

In each trial of each task, participants first saw the variable, economically less beneficial option followed by the constant, economically more beneficial but false or more effortful option. The payoff of the first options varied between 1 and 5 CHF whereas the second option always led to a fixed payoff of 5 CHF. Therefore the cost of choosing the first option varied between 0 and 4 CHF. Both task type and payoff-levels were randomly intermixed across trials during the experiment. Trials were separated from each other by an intertrial interval (ITI) with mean

duration of 4 s, varying from 2 to 6 s (Figure 1 A). In every trial the first event consisted of the presentation of a cue (1 s) shown at the top of the screen that indicated which kind of task participants had to perform. The cue ”Announce” referred to the truthtelling task, “Perform” to the valuation task and “Calculate” to the effort task. The first, variable option was then shown for 3 s in the center of the screen. In the truthtelling task the true earning (31 cents per share) was presented. In the valuation task the option for the low value project and in the effort task the option for solving one problem was shown. Below the option the CEO compensation for that option was indicated together with the corresponding participant payoff in parentheses. After an interstimulus interval consisting of a blank screen (mean of 4 s, varying between 2 and 6 s) the second, constant option was presented together with the first option. The second option was the false report (35 cents per share) in the truthtelling task, the high value project in the valuation task and the option for performing five calculations in the effort task.

Participants were asked to make up their mind already during the presentation of the first option. This was possible as only the first option varied across trials whereas the second option remained constant. The decision was conveyed during the presentation of the second option. To prevent motor preparation during presentation of the first option the two options were presented randomly either on the left or the right side of the screen during the response phase. Participants had 2 s to indicate their choice by performing a button press. When subjects pressed a button, the color of the written text on the screen changed from white to yellow to indicate that a response had been recorded. To keep them engaged throughout the task, in each trial in which participants failed to respond or responded too slowly, CHF 1 was deducted from their final monetary payment. These trials were repeated later. On average the percentage of trials missed by participants was 1.5% ± 0.6% (mean ± SEM; range 0 - 16%) with no significant difference between the tasks ($F_{(1,32)} = 0.35$, $p = 0.70$). Over the course of the experiment, each participant completed 75 truthtelling trials, 75 valuation trials and 75 effort trials. Each of the five payoff levels was presented 15 times per task. Stimulus presentation and response recording were controlled using Cogent 2000 (Wellcome Department of Imaging Neuroscience, London, UK) as implemented in Matlab.

Measurement of Protected Values

After scanning we assessed to which extent participants treated truthfulness as a protected value and felt committed to telling the truth as described previously (Tanner et al., 2009; Gibson et al., 2013). The questionnaire contained two subscales which measure protected values from two

different angles. The direct subscale assesses how reluctant participants are to sacrifice the value of telling the truth in the specific context of a hypothetical CEO's decisions regarding the reporting of earnings. For example participants indicate how strongly they agree with the statement that one should not sacrifice truthfulness, no matter what the (material or other) benefits are. By contrast, the indirect subscale examines how participants evaluate the decisions of others by assessing their subjective emotional reactions to violations of honesty by a hypothetical CEO who reports company earnings. For instance, participants indicate how blameworthy it is in their opinion when CEO's modify earnings reports. We were able to construct three indices of protected values, one based on the direct, the second based on the indirect subscale and the third based on the mean of the two subscales. The indices took a value between 0 (for an individual with no protected values) and 6 (for an individual with maximum protected values).

Behavioral analysis

The effect of protected values on choice behavior was analyzed using a series of logistic regression analyses. For each of the analyses, the dependent variable corresponded to the decision in a given trial. This was coded as a binary variable that took on the value of 1 if participants chose the honest (low effort or low value) option, and the value of 0 if participants chose the dishonest (high effort or high value) option. The independent variables (regressors) in the model included either cost-level, protected values or their interaction (each of the three protected value subscales – direct, indirect, and their mean – entered the models separately), while controlling for gender and age.

fMRI Data Acquisition

fMRI was performed on a Philips Achieva 3T whole-body scanner equipped with an eight-channel head coil (Philips Medical Systems, Best, The Netherlands) at the Laboratory for Social and Neural Systems Research, University of Zurich. We acquired gradient-echo T2*-weighted echoplanar images (EPIs) with blood-oxygen-level-dependent (BOLD) contrast (slices/volume, 37; repetition time, 2 s). Participants each completed five sessions of the experiment in the scanner, with short breaks between each session. 315-360 volumes were collected per session (the variation was due to individual differences in the number of repeated trials) along with 5 "dummy" volumes at the start of each session to allow for magnetization to stabilize to a steady state. Scan onset times varied randomly relative to stimulus onset times. Volumes were acquired

at a 15° tilt to the anterior commissure-posterior commissure line, rostral > caudal. Imaging parameters were the following: echo time, 30 ms; field-of-view, 220 mm; in-plane resolution, 2.75 mm; slice thickness, 3 mm; interslice gap, 0.5 mm. A T1-weighted structural image was also acquired for each participant. These high-resolution T1-weighted structural scans were coregistered to their mean EPIs and averaged to permit anatomical localization of the functional activations at the group level.

fMRI Data Analysis

fMRI data processing and statistical analyses were carried out using statistical parametric mapping (SPM8; Wellcome Department of Imaging Neuroscience, London, UK). Data preprocessing consisted of realignment, coregistration, segmentation, spatial normalization using the DARTEL toolbox and smoothing using a Gaussian kernel with a full width at half maximum of 10 mm.

The general linear model (GLM) identified brain regions in which activity correlated with cost-level. This GLM used the following set of regressors: Presentation of first Truthtelling option, Presentation of first Valuation option, and Presentation of first Effort option. All of these regressors were defined irrespective of actual decisions and modeled as stick functions with duration of 0s. For each of the regressors we included a parametric modulator capturing the variable cost-level. We also included regressors of no interest consisting of the onsets of decision time and the missed trials for each task. Finally, participant-specific movement parameters (three regressors for rotation and three for translation) were modeled, also as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. Main effects for cost of truthtelling, cost of valuation and cost of effort were computed on the single-subject level by performing separate t-tests for each parametric modulator. The resulting contrast images were taken up to the group-level where we used correlations with participant-specific percentage of truthtelling (or low value project/effort chosen) and protected values.

We used psycho-physiological interaction (PPI) analyses (Friston et al., 1997) to test for coupling differences due to variations in protected values. For this analysis we adapted the first GLM and computed a first-level model in which we defined separate regressors depending on cost. Specifically, for each of the tasks the model included a regressor that captured no (cost-level 0), low (cost-level 1 and 2) and high (cost-level 3 and 4) cost conditions. For each subject, the average time course was extracted from voxels in the dlPFC and dmPFC in which activity for cost of truthtelling correlated with percent truthtelling. With this time course as physiological

regressor, we performed a first-level model with the cost-level as a psychological regressor (high vs. low cost), and a psycho-physiological interaction (PPI) regressor. We then used the PPI regressor to perform a second-level correlation analysis with individual protected values.

Correction for multiple comparisons was performed in the whole brain at the cluster-level ($p < 0.05$, cluster-inducing threshold: $p < 0.005$). Moreover, to interrogate some of the findings in more detail we used small-volume correction in regions of interest that were identified in independent, preceding whole-brain-corrected analyses.

Results

Behavioral results

Choice behavior irrespective of protected values

Our experimental design aimed to create trade-offs between moral and economic motives. In particular, we chose a trade-off situation that CEOs actually face in real life. CEO compensation is frequently tied to stock price performance and thereby to the company earnings CEOs announce. In our truthtelling task participants assumed the role of a CEO and decided whether to announce true earnings, which would reduce their compensation. Alternatively they could falsely report higher earnings which would result in a higher compensation. By contrast to the truthtelling task, in our effort task participants faced a trade-off situation without a moral aspect. They decided whether to work less, which would reduce their compensation, or to work harder and achieve a higher compensation. Thus, there was a non-moral trade-off between working more for higher payment and working less for lower payment in this control task. Importantly, the economically more costly option (telling the truth, choosing to work less) involved the same variable compensation reductions (i.e. economic costs) in both tasks. In the valuation task participants decided whether to choose a low-value project that would reduce their compensation or a high-value project that would result in a higher compensation. There was no trade-off in this control task as there were no costs for choosing the high value option. However, the variable economic value difference between the two options was equivalent to that of the other two tasks. Thus, at equal economic stakes, one task involved a trade-off with moral values, one a trade-off with non-moral value and one involved no trade-off at all.

In the truthtelling task, participants on average told the truth in a bit less than half of the trials (39.8% ± 5.1% (mean ± SEM)), and showed a wide variation in their choices with some participants choosing to tell the truth in all trials and others doing so in very few trials (range 6.7 - 100%). In the effort task participants chose the low effort option with similar frequencies (40.1% ± 5.0%, range 0 - 93.3%) while in the valuation task they chose the low value option less often (18.8% ± 2.6%, range 0 - 60%). Moreover, across participants, the percentage of truthtelling was correlated with the percentage of low effort chosen ($R^2 = 0.29$, $p < 0.01$). The truthtelling and the valuation task ($R^2 = 0.04$, $p = 0.26$) and the effort and the valuation task ($R^2 = 0.02$, $p = 0.47$) did not show a correlation.

We tested whether participants' choices differed as the economic costs of truthfulness changed by regressing the choices against the cost-level. This revealed significant effects of cost ($\beta = -0.95 \pm 0.13$ (mean ± SEM), $t = -7.29$, $p < 0.001$) on truthful decisions (fewer truthful decisions with increasing economic costs), suggesting that at least some participants traded off the economic cost of telling the truth with its moral benefits. We observed similar effects for the two control tasks: the cost-level predicted whether or not participants chose the low value/effort option (valuation task: $\beta = -1.50 \pm 0.19$, $t = -8.06$; effort task: $\beta = -1.02 \pm 0.15$, $t = -6.91$; both $p < 0.001$).

Protected values and their impact on choice behavior

Participants exhibited individual differences in the direct measure of protected values, the indirect measure and the mean of the two subscales (direct measure: 4.1 ± 0.2 (mean ± SEM), range 1.8 – 6.0; indirect measure: 4.5 ± 0.2, range 2.4 – 6.0; mean score: 4.3 ± 0.2, range 2.9 – 6.0). These values were similar to those reported in a previous study (direct measure: 3.35 ± 1.29; indirect measure: 4.2 ± 1.1; mean score: 3.8 ± 1.03; see Gibson et al., 2013). Moreover, as in the previous study, the direct and the indirect measures were correlated ($R^2 = 0.23$, $p < 0.01$).

Without taking cost level into account, logistic regressions revealed no significant relations between any measure of protected values and choice behavior in any of the tasks (all $\beta < 0.25$, $t < 1.70$, $p > 0.09$). However, higher protected values should allow participants to resist the temptation of sacrificing moral values for economic benefits particularly when the economic costs for telling the truth are high. Thus, high protected values should lead to more honest decisions in higher cost conditions. We assessed this hypothesis in an interaction analysis using the direct and indirect subscales separately and together based on the means of the two subscales.

In agreement with the hypothesis, we observed a positive interaction between direct protected values and economic costs of truthfulness ($\beta = 0.28 \pm 0.13$, $t = 2.15$, $p < 0.05$). In other words, when economic costs of truthfulness were high participants with stronger protected values chose the truthful option more often compared to participants with weaker protected values. By contrast, no significant interactions between protected values and economic costs of truthfulness arose for the indirect ($\beta = -0.10 \pm 0.13$, $t = -0.76$, $p = 0.45$) and the total score measure ($\beta = 0.09 \pm 0.14$, $t = 0.61$, $p = 0.54$). Thus, particularly the direct measure of protected values predicted whether or not individuals would trade off truthfulness against the economic cost as the costs increased (Figure 1 B). We therefore primarily used this subscale for the analysis of the imaging data.

Importantly, in the valuation and effort task, we found no interaction effect induced by participants higher in direct protected values choosing the more costly option more often at higher costs. While there was no interaction effect for the effort task ($\beta = 0.12 \pm 0.13$, $t = 0.92$, $p = 0.36$), in the valuation task participants with high protected values chose the low value option less often in the zero cost condition ($\beta = 0.31 \pm 0.14$, $t = 2.16$, $p < 0.05$). Note though that this interaction effect is quite different from that in the truthtelling task, where participants with high protected values chose the truthful option more often in high cost conditions. Indeed, when excluding the zero cost condition (and thus including only trade-off conditions, where there was an actual economic cost for telling the truth) the interaction effect for the valuation task disappeared completely ($\beta = 0.14 \pm 0.16$, $t = 0.88$, $p = 0.38$) but was still present, at least at trend level, for the truthtelling task ($\beta = 0.30 \pm 0.16$, $t = 1.91$, $p = 0.06$). Taken together, these data suggest that high protected values reduced participants' willingness to trade off specifically moral as opposed to no-moral values against economic costs.

In further regression analyses we also tested whether response times were related to protected values and cost-levels. We observed neither a main effect of cost ($\beta = 7.49 \pm 5.07$, $t = 1.48$, $p = 0.15$) or protected values ($\beta = 13.56 \pm 15.18$, $t = 0.89$, $p = 0.38$) nor a cost x protected values interaction ($\beta = 1.73 \pm 5.79$, $t = 0.30$, $p = 0.77$). Moreover, we also found no effect of choice on response times ($\beta = -17.77 \pm 30.79$, $t = -0.58$, $p = 0.57$).

Neuroimaging results

DLPFC and DMPFC code truthtelling costs with increasing propensity to be honest

In a first step, we aimed to find a basic relation between coding the cost of telling the truth and the observed behavioral propensity to tell the truth, irrespective of protected values. We expected that cost would be differentially encoded depending on how often participants show honest behavior. We therefore searched for brain areas where activation related to cost-level (parametric modulator) changed as a function of percent truthtelling. As a trade-off between moral and economic incentives arises only when at least some economic costs are incurred for telling the truth we performed this analysis without the zero cost condition. We found that neural coding of the cost of truthtelling increased with the individual percentage of truthtelling in the dlPFC and dmPFC (Figure 2A and B; dlPFC: -30, 56, 26; $t_{(30)} = 4.92$; Figure 2C and D; dmPFC: -6, 16, 44; $t_{(30)} = 4.48$; both $p < 0.05$, whole-brain FWE cluster-level corrected). In other words, with increasing cost of telling the truth participants show stronger activation in the dlPFC and dmPFC the more often they make honest decisions.

We wondered whether the correlation of cost-related activity with the percentage of truthtelling would still emerge if we also accounted for individual differences in protected values. We therefore included protected values in the correlation analysis as a second covariate. Again we found stronger dlPFC and dmPFC activation with increasing cost and increasing percentage of truthtelling (dlPFC: -30, 56, 26; $t_{(29)} = 5.33$; dmPFC -6, 20, 40; $t_{(29)} = 5.19$; $p < 0.05$, both $p < 0.05$, whole-brain FWE cluster-level corrected).

We next tested whether the correlations between cost-related activation and percent truthtelling in the dlPFC and dmPFC are specific for moral as opposed to non-moral decisions by comparing the truthtelling task with the two control tasks. The analysis was based on the fact that in all tasks participants were confronted with the same economic cost-levels when choosing between the truth, low value or low effort option on the one hand and the lie, high value or high effort option on the other hand. Thus, the only additional component in the truthtelling task was the moral nature of the decision made. We therefore performed a second-level correlation analysis in which we tested for stronger correlation of truthtelling cost-related activation with percent truthtelling than of valuation cost-related activation with percent low value option chosen and of effort cost-related activation with percent low effort option chosen. We observed no significant activation specific for the truthtelling task. This finding is in line with the notion that cost coding in the dlPFC and dmPFC is related to the individual propensity of incurring

costs more generally, rather than specifically to the propensity of incurring economic costs for moral benefits.

DLPFC-IFG and DMPFC-IFG interactions increase with protected values

The analysis above identified differential cost coding in the dmPFC and dlPFC in relation to individual differences in behavior rather than protected values. However, our behavioral analyses indicated that higher protected values support honest decisions as the cost of telling the truth increases. In a next step we therefore asked how protected values affect the brain specifically in high cost situations. In particular we investigated whether the dlPFC and dmPFC that code costs as a function of percent truthtelling change their interaction with other brain regions with respect to protected values. We therefore formed two first-level models with activity either from the dlPFC or the dmPFC region identified above as physiological regressor, the cost-level as a psychological regressor (high vs. low cost), and a psycho-physiological interaction (PPI) regressor corresponding to the product of the first two regressors. The PPI regressor then served for a second-level correlation analysis with protected values. We found that functional connectivity between the dlPFC and the IFG as well as between the dmPFC and the IFG differed significantly as a function of cost-level and protected values (Figure 3A; dlPFC: -46, 28, 30; $t_{(30)} = 4.56$; Figure 3C; dmPFC: -44, 28, 28; $t_{(30)} = 5.12$; $p < 0.05$, both whole-brain FWE cluster-level corrected; additionally, whole-brain cluster-level corrected activation was found in the parietal cortex (-46, -32, 38; $t_{(30)} = 4.29$)). Specifically, both regions showed stronger functional connectivity with the IFG in high cost conditions compared to low cost conditions as protected values increased (Figure 3B and D).

We wondered whether the modulation of dlPFC-IFG and dmPFC-IFG connectivities by protected values would still emerge when we also accounted for individual differences in the percentage of truthtelling. We therefore included a second regressor with percent truthtelling as a covariate of no interest in the correlation analysis and again found that both dlPFC-IFG and dmPFC-IFG coupling increased more in high than low cost conditions as protected values increased (dlPFC-IFG: -46, 28, 30; $t_{(29)} = 4.86$; dmPFC-IFG: -44,28,28; $t_{(29)} = 5.40$, both $p < 0.05$, whole-brain FWE cluster-level corrected).

We next tested whether the modulation of the dlPFC-IFG and dmPFC-IFG connectivity by protected values is specific for the truthtelling task by comparing the activity pattern with the other two tasks. We therefore included the two control tasks in the PPI model with the cost-levels (high vs. low) of the valuation and effort task respectively as psychological regressor and

correlated the strength of connectivity with the individual differences in protected values for all tasks. We found that the cost-level dependent connectivity between dlPFC and IFG related more strongly to protected values in the truthtelling task as compared to the other two tasks (Figure 4; -46, 24, 16; $t_{(60)} = 4.52$; $p < 0.05$, whole-brain FWE cluster-level corrected). By contrast, the cost- and protected-value-related increase of dmPFC connectivity with the IFG was not specific for the truthtelling task, ($p > 0.8$). Thus, protected values appear to exert their effects on moral decision making particularly via a connection between the dlPFC and the IFG.

Discussion

In the present study we investigated how individual differences in behavior and protected values are represented in the brain particularly during moral decisions involving honesty. We demonstrate that the more individuals actually behave honestly the more the dlPFC and dmPFC are engaged by increasing economic costs of honesty. Importantly, in high compared to low cost situations, both the dlPFC and dmPFC show stronger coupling with the IFG with increasing levels of protected values. Moreover, the relation of dlPFC-IFG connectivity with individual differences in protected values is specific for moral decisions. These findings provide novel insights into how prefrontal connectivity underpins individual variability in a personality trait that promotes honesty.

In line with previous behavioral research (Gibson et al., 2013) we find that personal factors (protected values) and situational factors (cost-level) predict whether individuals behave honestly. More specifically, when economic costs of truthfulness were high participants with stronger protected values chose the truthful option more often compared to participants with weaker protected values. This suggests that protected values are instrumental in driving truthful decisions particularly when the costs of truthfulness are higher. It should be noted, however, that in the current study we only find an effect with the direct measure of protected values. By contrast, in a previous study both the direct and indirect subscales showed a relation to behavior (Gibson et al., 2013). The direct measure captures the more cognitive dimension of protected values (Baron and Spranca, 1997), whereas the indirect measure emphasizes the emotional aspects and affective reactions individuals experience when they evaluate dishonest behavior (Tetlock et al., 2000). One obvious difference between our current study and the one conducted previously is that our participants completed many trials, whereas in the previous study they

decided only once for each cost-level. Thus, one explanation could be that it was not possible to capture the affective aspect with our design as emotional responses wear off more quickly than cognitive responses. Moreover, it should be noted that in contrast to the previous study both the original (effort) and the added (valuation) control tasks were presented randomly intermixed with the truthtelling task to prevent non-specific block effects. Yet, confronting participants with moral decisions in isolation could have enhanced the affective responses to the proposal of dishonesty for personal gain in the previous study. Thus, intermixing of tasks in our study could have weakened emotional reactions to the moral aspects of the truthfulness task.

In our study, cost coding in the dlPFC and the dmPFC was related to individual differences in honesty. Previous research found especially the dlPFC, but also the dmPFC to play a crucial role during honest and dishonest behavior. For example, the dlPFC and dmPFC have been associated with deception during instructed and spontaneous lying (Sip et al., 2008; Abe, 2009, 2011), which has been taken as evidence that lying is more demanding and thus needs more cognitive control than telling the truth. However, the dlPFC and dmPFC also have been associated with honesty. Two recent studies found that individuals who are more prone to economic gain and exhibit high levels of dishonesty show greater dlPFC (Greene and Paxton, 2009; Abe and Greene, 2014) and dmPFC (Greene and Paxton, 2009) engagement when making honest decisions. Our findings suggest a potential reconciliation of these discrepant findings. In particular, the dlPFC and dmPFC activity is driven by the combination of the individuals' propensity to be honest and how costly honesty actually is.

The increased dlPFC and dmPFC responses and their connectivity with the IFG might represent an active control mechanism that implements honest behavior and biases honest individuals to refrain from dishonest behavior specifically at higher economic cost. These regions have been consistently implicated in cognitive control and response inhibition (Botvinick et al., 2001; Miller and Cohen, 2001; Aron, 2007; Carter and van Veen, 2007). We note that this interpretation is different to that proposed by previous studies (Sip et al., 2008; Abe, 2009, 2011) which suggest that deceptive rather than honest behavior engages the control network. The present study differs in several ways from the ones conducted previously. First, in a recent study that has investigated spontaneous lying (Greene and Paxton, 2009) participants were given the opportunity to behave dishonestly by lying about the accuracy of their predictions about the outcomes of computerized coin flips. Thus, in lie trials participants had to override their previous prediction, which may more difficult than the lie decisions in our study. Second, in our study participants were explicitly told that in real life CEOs could report false earnings

within accounting laws. Consequently, in our experiment the default action was more likely to be the dishonest option. Thus, whether honesty or dishonesty requires the engagement of control regions might depend on which of the two represents the default action.

It has been proposed that protected values derive from rules that prohibit certain actions (Baron and Spranca, 1997). On a very basic level these rules might be stored as semantic knowledge, which is retrieved when individuals are faced with decision situations that are associated with those rules. Interestingly, the IFG has been associated with semantic rule retrieval and processing (Bunge, 2004; Badre et al., 2005; Souza et al., 2009). Thus, the IFG could provide the dlPFC and dmPFC with input that represents semantic rules which biases participants towards honesty irrespective of consequences (although it should be kept in mind that directionality cannot be inferred from PPI effects). In agreement with this interpretation, a recent study has shown responses in the IFG when individuals refused to accept money for changing previously reported moral views (Berns et al., 2012). Our results reveal that the connectivity in individuals with strong protected values is enhanced specifically when high economic costs are involved and thus strong control mechanisms are needed to protect the moral values.

Interestingly, individuals with low protected values exhibited increased dlPFC-IFG and dmPFC-IFG coupling during low cost conditions. Given the proposed role of the IFG in implementing honest behavior it is conceivable that the coupling biases individuals with weak protected values towards honesty only in situations where the costs for telling the truth are low. This interpretation would imply that the connectivity itself is not specific to individuals with strong protected values but that it is specifically engaged in strong protected value individuals when the economic cost and thus the temptation to switch to dishonest behavior is high. Thus, in all individuals the dlPFC-IFG and dmPFC-IFG coupling translates into increased honest behavior, but only individuals with strong protected values exhibit stronger coupling during high cost situations which shields their moral values against economic considerations.

Our results re-visit the central question of whether moral decisions are represented by domain-specific or domain-general neural mechanisms (Greene and Haidt, 2002; Hauser, 2006). Our findings suggest that the correlation between cost coding in the dlPFC and dmPFC with individual honesty is not specific to moral trade-off situations, but represents a more general relation of coding the costs of any decision to individual differences in how these costs influence the decision. By contrast, individual differences in protected values predict the strength of dlPFC-IFG connectivity specifically in the honesty task, but not in the two control tasks. Thus,

individual differences that arise from reluctance to consider economic consequences of moral behavior seem to elicit specific connectivity patterns that differ from those associated with individual differences in decisions involving trade-offs between consequences. These findings imply that the brain regions involved in moral decision making are not specific to moral decisions, but the way they interact with each other as a function of cost and protected values is specifically related to moral aspects of decisions.

The dlPFC-IFG connectivity but not the dmPFC-IFG connectivity showed specificity for protected values when compared with the non-moral control tasks. It has been proposed that the dmPFC plays a role in monitoring and specifying cognitive control, while the dlPFC is assumed to be responsible for the regulative function of control, that is for actually implementing the control signal to execute the control-demanding action (MacDonald et al., 2000; Botvinick et al., 2004; Botvinick, 2007; Carter and van Veen, 2007; Shenhav et al., 2013). Thus, we speculate that in our task the monitoring function represented by dmPFC-IFG connectivity relates to protected values also in the control tasks. By contrast, the implementation function represented by dlPFC-IFG connectivity relates to protected values only in the truthtelling but not in the other tasks, as only in the truthtelling task protected values predict how economic costs are processed. Thus, it might be that conflict monitoring driven by protected values partly also affects other domains of conflict, not only those involving moral aspects. By contrast, the implementation of control as mediated by protected values dependent dlPFC-IFG coupling plays a role specifically during moral decisions. By extension, personality-related moral cognition could be both domain-general and domain-specific depending on the specific mental function that is investigated.

The interpretation mentioned above also provides an understanding on protected values more generally by suggesting that protected values may be tied to self-regulatory processes and exert their influence on behavior via control mechanisms. Interestingly, recent studies have shown that active self-control indeed supports honest behavior (Mead et al., 2009; Gino et al., 2011; Shalvi et al., 2012). Taken together, our findings suggest that stronger protected values could be linked to stronger self-control mechanisms, a possibility that may warrant future research.

In conclusion, our results provide the first evidence that neural connectivity patterns during decisions involving honesty are modulated by individual differences in protected values. Individual variability in actual behavior relates to parametric cost coding in the dlPFC and dmPFC, and both of these relations occurred also during more general, non-moral decisions. By

contrast, individual differences in protected values modulate the dlPFC-IFG connectivity specifically during moral decision making. Our results thereby highlight the role of individual differences in moral attitudes and behavior and provide an explanation why some people decide to follow a moral principle whereas others do not.

References

Abe N (2009) The neurobiology of deception: evidence from neuroimaging and loss-of-function studies. Curr Opin Neurol 22:594–600.

Abe N (2011) How the brain shapes deception: an integrated review of the literature. Neuroscientist 17:560–574.

Abe N, Greene JD (2014) Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. J Neurosci 34:10564–10572.

Abe N, Suzuki M, Tsukiura T, Mori E, Yamaguchi K, Itoh M, Fujii T (2006) Dissociable roles of prefrontal and anterior cingulate cortices in deception. Cereb Cortex 16:192–199.

Aron AR (2007) The neural basis of inhibition in cognitive control. Neuroscientist 13:214–228.

Badre D, Poldrack RA, Paré-Blagoev EJ, Insler RZ, Wagner AD (2005) Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. Neuron 47:907–918.

Baron J, Spranca M (1997) Protected values. Organ Behav Hum Dec 70:1–16.

Baumgartner T, Fischbacher U, Feierabend A, Lutz K, Fehr E (2009) The neural circuitry of a broken promise. Neuron 64:756–770.

Berns GS, Bell E, Capra CM, Prietula MJ, Moore S, Anderson B, Ginges J, Atran S (2012) The price of your soul: neural evidence for the non-utilitarian representation of sacred values. Philos Trans R Soc Lond B Biol Sci 367:754–762.

Botvinick MM (2007) Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. Cogn Affect Behav Neurosci 7:356–366.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624–652.

Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci 8:539–546.

Bunge SA (2004) How we use rules to select actions: a review of evidence from cognitive neuroscience. Cogn Affect Behav Neurosci 4:564–579.

Carter CS, van Veen V (2007) Anterior cingulate cortex and conflict detection: an update of theory and data. Cogn Affect Behav Neurosci 7:367–379.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

Ganis G, Kosslyn SM, Stose S, Thompson WL, Yurgelun-Todd DA (2003) Neural correlates of different types of deception: an fMRI investigation. Cereb Cortex 13:830–836.

Gibson R, Tanner C, Wagner A (2013) Preferences for truthfulness: heterogeneity among and within individuals. Am Econ Rev 103:532–548.

Gino F, Schweitzer ME, Mead NL, Ariely D (2011) Unable to resist temptation: How self-control depletion promotes unethical behavior. Organ Behav Hum Dec 115:191–203.

Greene JD, Paxton JM (2009) Patterns of neural activity associated with honest and dishonest moral decisions. Proc Natl Acad Sci USA 106:12506–12511.

Greene J, Haidt J (2002) How (and where) does moral judgment work? Trends Cogn Sci 6:517–523.

Hauser MD (2006) The liver and the moral organ. Soc Cogn Affect Neurosci 1:214–220.

Langleben DD, Loughead JW, Bilker WB, Ruparel K, Childress AR, Busch SI, Gur RC (2005) Telling truth from lie in individual subjects with fast event-related fMRI. Hum Brain Mapp 26:262–272.

MacDonald AW, Cohen JD, Stenger VA, Carter CS (2000) Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science 288:1835–1838.

Mead NL, Baumeister RF, Gino F, Schweitzer ME, Ariely D (2009) Too tired to tell the truth: self-control resource depletion and dishonesty. J Exp Soc Psychol 45:594–597.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Shalvi S, Eldar O, Bereby-Meyer Y (2012) Honesty requires time (and lack of justifications). Psychol Sci 23:1264–1270.

Shenhav A, Botvinick MM, Cohen JD (2013) The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79:217–240.

Shenhav A, Greene JD (2010) Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron 67:667–677.

Sip KE, Lynge M, Wallentin M, McGregor WB, Frith CD, Roepstorff A (2010) The production and detection of deception in an interactive game. Neuropsychologia 48:3619–3626.

Sip KE, Roepstorff A, McGregor W, Frith CD (2008) Detecting deception: the scope and limits. Trends Cogn Sci 12:48–53.

Souza MJ, Donohue SE, Bunge SA (2009) Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. Neuroimage 46:299–307.

Spence SA, Farrow TF, Herford AE, Wilkinson ID, Zheng Y, Woodruff PW (2001) Behavioural and functional anatomical correlates of deception in humans. Neuroreport 12:2849–2853.

Tanner C, Ryf B, Hanselmann M (2009) Geschützte Werte Skala (GWS): Konstruktion und Validierung eines Messinstrumentes (Protected values measure: construction and first validation of an instrument to assess protected values). Diagnostica 55:174–183.

Tetlock PE, Kristel OV, Elson SB, Green MC, Lerner JS (2000) The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. J Pers Soc Psychol 78:853–870.

Tobler PN, Kalis A, Kalenscher T (2008) The role of moral utility in decision making: An interdisciplinary framework. Cogn Affect Behav Neurosci 8:390–401.

Figure legends

Fig. 1 Experimental design and behavioral results. **A**, In each trial of each task, participants first viewed a fixation cross for a variable ITI of 2 – 6 s followed by the presentation of a cue (1 s) that indicated which kind of task participants had to perform. The first, variable option was then shown for 3 s. In the truthtelling task the true earning (31 cents per share) was presented, in the valuation task the option for the low value project (project ZEM) and in the effort task the option for solving one problem was shown. The payoff of the first options varied between 1 and 5 CHF. Below the option the CEO compensation was indicated together with the corresponding participant payoff in parentheses. After an interstimulus interval (2 - 6 s) the second, constant option was presented together with the first option. The second option was the false report (35 cents per share) in the truthtelling task, the high value project in the valuation task (project XIR) and the option for performing five calculations in the effort task. Upon presentation of the second option participants had 2 s to indicate their choice by performing a button press. When subjects pressed a button, the color of the written text on the screen changed from white to yellow to indicate that a response had been recorded. **B**, The difference in the percentage of truthtelling increased as a function of cost and protected values. When economic costs of truthfulness were high participants with stronger protected values (upper tercile) were more honest compared to those with low protected values (lower tercile), suggesting that protected values are more important in determining truthful decisions when the costs of truthfulness are higher. Error bars indicate SEMs.

Fig. 2 Activity in DLPFC and DMPFC reflects truthtelling costs with increasing propensity to be honest. **A and B**, DLPFC responses to the cost of truthtelling increased with the individual percentage of truthtelling (-30, 56, 26; $t_{(30)} = 4.92$; $p < 0.05$, whole-brain FWE cluster-level

corrected). **C and D**, DMPFC responses to the cost of truthtelling increased with the individual percentage of truthtelling (-6, 16, 44; $t_{(30)} = 4.48$; $p < 0.05$, whole-brain FWE cluster-level corrected). Color bars indicate z-scores.

Fig. 3 DLPFC-IFG and DMPFC-IFG interactions increase with truthtelling costs and protected values. **A**, Functional connectivity between the DLPFC (seed shown as inset) and the IFG was increased in high compared to low cost conditions as protected values increased (-46, 28, 30; $t_{(30)} = 4.56$; whole-brain FWE cluster-level corrected). Color bar indicates z-score. **B**, Connectivity strength between the DLPFC and IFG plotted separately for participants with high and low protected values (PV) and for high and low cost conditions. Error bars indicate SEMs. **C**, Functional connectivity between the DMPFC (seed shown as inset) and the IFG was increased in high compared to low cost conditions as protected values increased -44, 28, 28; $t_{(30)} = 5.12$; whole-brain FWE cluster-level corrected). Color bar indicates z-score. **D**, Connectivity strength between the DMPFC and IFG plotted separately for participants with high and low protected values (PV) and for high and low cost conditions. Error bars indicate SEMs.

Fig. 4 The DLPFC-IFG interaction pattern is specific for the truthtelling task. The increased connectivity between the DLPFC (inset) and the IFG for high compared to low cost conditions related more strongly to protected values in the truthtelling task as compared to the two control tasks (-46, 24, 16; $t_{(60)} = 4.52$; $p < 0.05$, whole-brain FWE cluster-level corrected). Color bar indicates z-score.

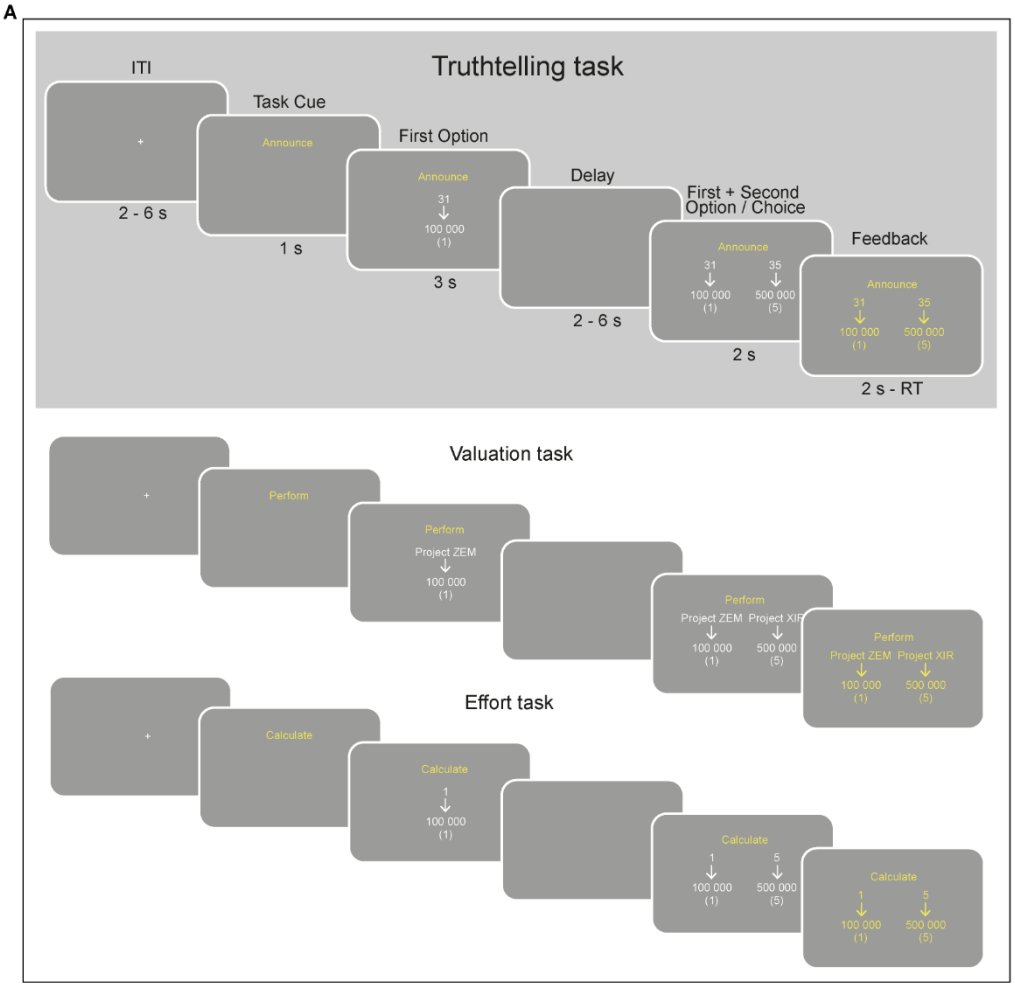
A
Truthtelling task
ITI
Task Cue
First Option
Delay
First + Second Option / Choice
Feedback
2 - 6 s
1 s
3 s
2 - 6 s
2 s
2 s - RT
Valuation task
Perform
Project ZEM
Project XIR
100 000
500 000
Effort task
Calculate
B
Percent truthtelling [%]
Cost of truthtelling (CHF)
high protected values
low protected values
0
20
40
60
80
100
0
1
2
3
4

Figure 1

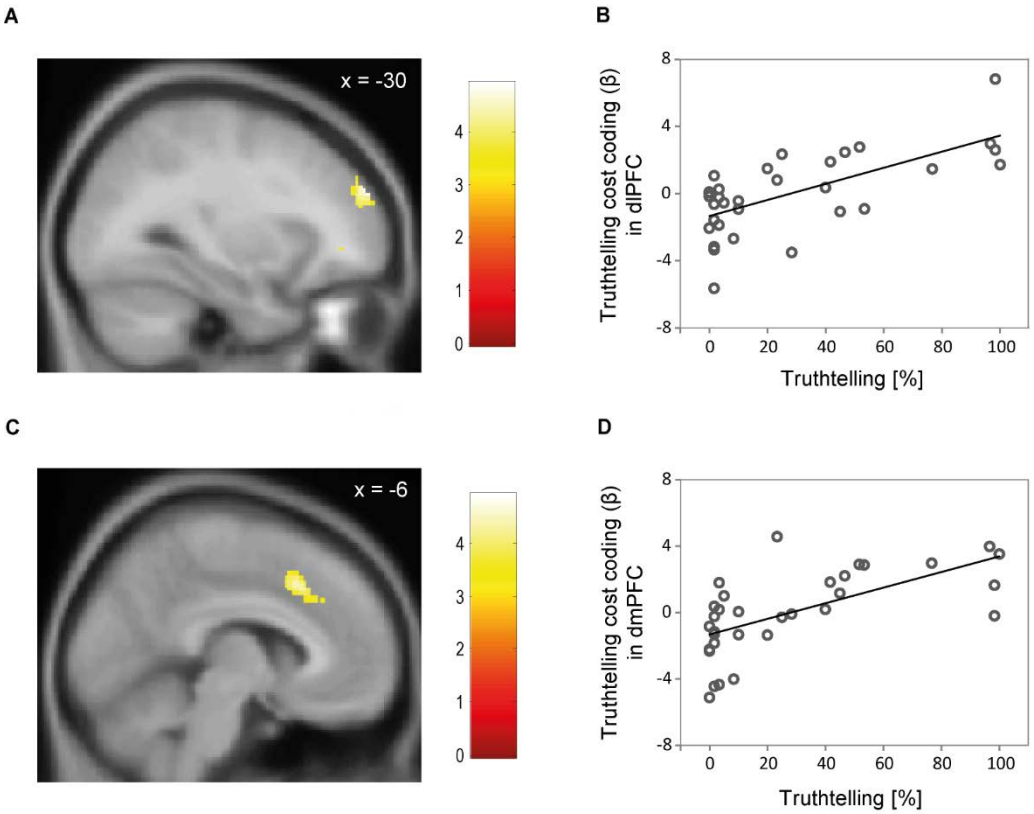
A
x = -30
4
3
2
1
0
B
Truthtelling cost coding (β) in dlPFC
8
4
0
-4
-8
0
20
40
60
80
100
Truthtelling [%]
C
x = -6
4
3
2
1
0
D
Truthtelling cost coding (β) in dmPFC
8
4
0
-4
-8
0
20
40
60
80
100
Truthtelling [%]

Figure 2

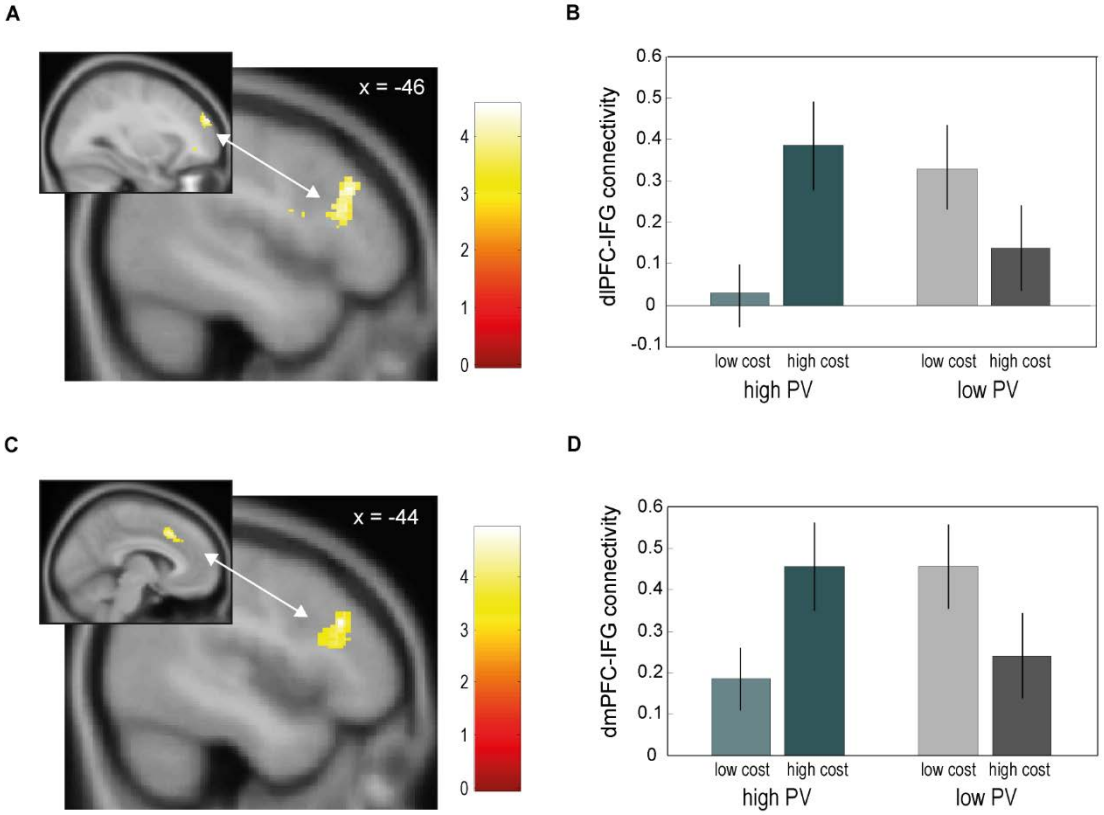

Figure 3

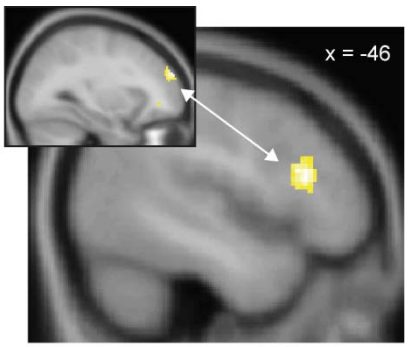

Figure 4

Curriculum Vitae

Azade Seid-Fatemi

Personal Data

Date of Birth:	May 20, 1983
Place of Birth:	Tehran, Iran
Nationality:	German

Education and Work Experience

10/2010 - 02/2015	Doctoral Studies in Neuroeconomics, Department of Economics, University of Zurich, Switzerland
05/2010 - 09/2010	Research associate, Institute of Cognitive Neuroscience, Ruhr-University Bochum, Germany
04/2007 - 04/2010	Biology studies (M.Sc.), University of Bochum, Germany
10/2003 - 11/2006	Biology studies (B.Sc.), University of Cologne, Germany
1994 - 2003	Bert-Brecht Gymnasium, Dortmund, Germany
	High school diploma (Abitur)